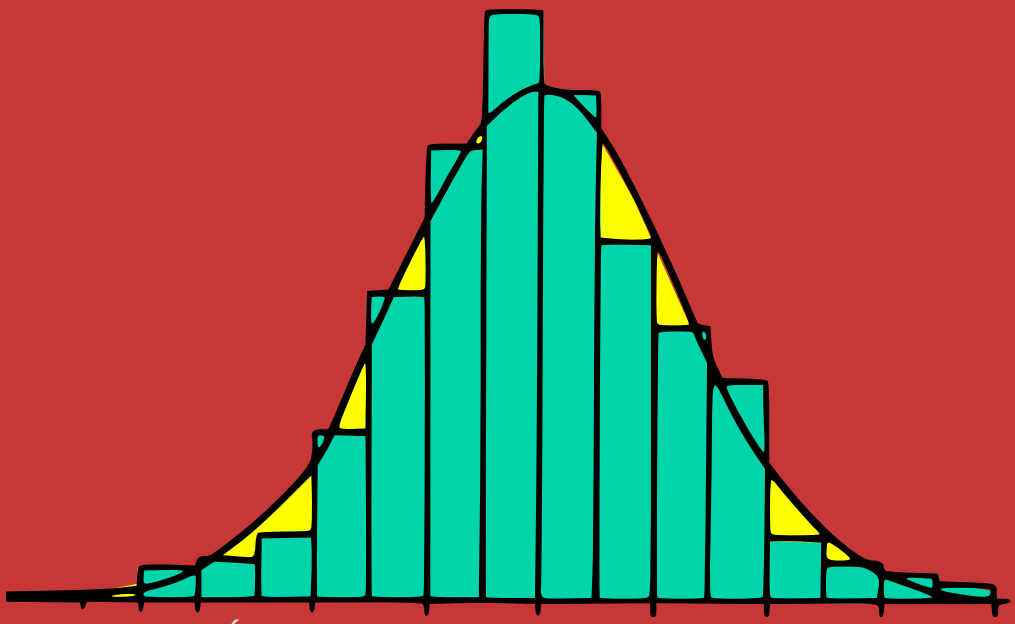
S. Aïvazian, I. Enukov, L. Mechalkine

ÉLÉMENTS DE MODÉLISATION ET TRAITEMENT PRIMAIRE DES DONNÉES



Éditions Mir Moscou

С. А. АЙВАЗЯН,
И. С. ЕНЮКОВ,
Л. Д. МЕШАЛКИН

ПРИКЛАДНАЯ СТАТИСТИКА

ОСНОВЫ МОДЕЛИРОВАНИЯ И ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ

ИЗДАТЕЛЬСТВО «ФИНАНСЫ И СТАТИСТИКА»
МОСКВА

S. AÏVAZIAN, I. ÉNUKOV, L. MÉCHALKINE

ÉLÉMENTS DE MODÉLISATION ET TRAITEMENT PRIMAIRE DES DONNÉES

ÉDITIONS MIR • MOSCOU

Traduit du russe
par DJILALI EMBAREK

На французском языке

© Издательство «Финансы и статистика», 1983
© traduction française, Editions Mir, 1986

TABLE DES MATIÈRES

Avant-propos 12

PREMIÈRE PARTIE. STATISTIQUE APPLIQUÉE: ESSENCE ET OBJECTIF (principes méthodiques généraux) 15

CHAPITRE PREMIER. La statistique appliquée: une discipline scientifique autonome 15

1.1. Lien entre la statistique appliquée et les autres disciplines statistiques et principales étapes de la recherche statistique 15

1.1.1. Définition de la statistique appliquée 15

1.1.2. Deux façons d'interpréter les données initiales et deux approches de leur traitement statistique 16

1.1.3. Principales étapes du traitement statistique des données initiales 20

1.2. Formulation d'optimisation des principaux problèmes de statistique appliquée et stabilité des inférences statistiques 27

1.2.1. Lien entre la formulation d'optimisation des principaux problèmes de statistique appliquée et la stabilité des inférences statistiques 27

1.2.2. Etude statistique des dépendances des variables . . 28

1.2.3. Classification des objets ou des variables 31

1.2.4. Abaissement de la dimension de l'espace factoriel et sélection des variables les plus informatives . . . 32

Conclusions 33

CHAPITRE 2. Méthode de raisonnement probabiliste en statistique appliquée 35

2.1. Théorie des probabilités et conditions de son application . . 35

2.1.1. Ensemble statistique et « jeu du hasard » 35

2.1.2. Théorie des probabilités et conditions d'ensemble statistique 36

2.1.3. Classification des situations réelles par les conditions d'ensemble statistique 37

2.2. « Rapports » entre la théorie des probabilités et la statistique mathématique 40

2.2.1. Mode statistique de prise de décision 40

2.2.2. Mode probabiliste de raisonnement 40

2.2.3. Mode probabilisto-statistique (ou mathématico-statistique) de prise de décision 41

Conclusions 43

CHAPITRE 3. Modèles mathématiques en statistique appliquée . . . 45

3.1. A quoi servent les modèles mathématiques? 45

3.1.1. Deux façons de voir la modélisation statistique . . . 45

3.1.2. Notion de modèle mathématique 46

3.2. Schéma logique général et principales étapes de modélisation mathématique 48
3.2.1. Principales étapes de modélisation 48
3.2.2. Modélisation du mécanisme d'un événement au lieu d'une photographie statistique formelle 50
3.3. Notion de simulation statistique 52
3.4. Objections contre les modèles mathématiques 53
3.5. Types de modèles mathématiques les plus souvent utilisés en statistique appliquée 54
3.5.1. Modèles des lois de probabilité de variables aléatoires 54
3.5.2. Modèles probabilistes linéaires 54
3.5.3. Généralisation des modèles linéaires 56
3.5.4. Modèles géométriques 57
3.5.5. Modèles markoviens 58
Conclusions 58

DEUXIÈME PARTIE. FONDEMENTS MATHÉMATIQUES DE LA THÉORIE DES PROBABILITÉS 60

CHAPITRE 4. Opérations sur les événements aléatoires et les probabilités de leur réalisation 60

4.1. Espace probabilisé discret 60
4.1.1. Expérience aléatoire 60
4.1.2. Evénements aléatoires et opérations sur eux 61
4.1.3. Espace probabilisé. Probabilités et opérations sur elles 65
4.2. Espace probabilisé continu (axiomatique de Kolmogorov) 74
4.2.1. Spécificité du cas général (continu) de l'espace probabilisé 74
4.2.2. Evénements aléatoires, probabilités et opérations (approche axiomatique de Kolmogorov) 76
Conclusions 79

CHAPITRE 5. Variables aléatoires 81

5.1. Définition et exemples de variables aléatoires 81
5.2. Valeurs possibles et observées d'une variable aléatoire . . . 83
5.3. Types de variables aléatoires 83
5.4. Loi de probabilité d'une variable aléatoire. Population générale et échantillon 86
5.4.1. Loi de probabilité 86
5.4.2. Population générale et échantillon 90
5.4.3. Principales méthodes d'organisation de l'échantillonnage 93
5.5. Méthodes de définition de la loi de probabilité: fonction de répartition, densité de probabilité et leurs analogues empiriques 95
5.5.1. Fonction de répartition d'une variable aléatoire à une dimension 95
5.5.2. Densité de probabilité d'une variable aléatoire à une dimension 98
5.5.3. Fonction de répartition et densité de probabilité conjointes. Indépendance statistique des variables aléatoires 101
5.6. Principales caractéristiques numériques des variables aléatoires et leurs analogues empiriques 106
5.6.1. Notion d'espérance mathématique et de moment . . 107

5.6.2. Caractéristiques du centre de groupement des valeurs de la variable aléatoire 109
5.6.3. Caractéristiques du degré de dispersion d'une variable aléatoire 112
5.6.4. Série variationnelle et statistiques de rang 114
5.6.5. Quantiles et points de pourcentage d'une répartition 117
5.6.6. Dissymétrie et aplatissement 118
5.6.7. Principales caractéristiques des répartitions multidimensionnelles (covariance, corrélation, variance généralisée, etc.) 120
Conclusions 123

CHAPITRE 6. Les lois de probabilité modèles les plus couramment utilisées dans les recherches statistiques 126

6.1. Lois de probabilité servant à décrire les mécanismes de processus ou de systèmes réels 126
6.1.1. Lois décrivant une suite d'épreuves de Bernoulli: la loi binomiale et la loi binomiale négative . . . 126
6.1.2. Loi hypergéométrique 129
6.1.3. Loi de Poisson 131
6.1.4. Loi polynomiale 132
6.1.5. Loi normale (de Gauss) 133
6.1.6. Loi lognormale 137
6.1.7. Loi uniforme (rectangulaire) 141
6.1.8. Loi de Weibull et loi exponentielle 143
6.1.9. Loi de Pareto 147
6.1.10. Loi de Cauchy 147
6.1.11. Quelques combinaisons des principales lois de probabilité modèles utilisées en statistique appliquée 148
6.2. Lois de probabilité utilisées dans les calculs statistiques . . 150
6.2.1. Loi du χ^2 150
6.2.2. Loi t de Student 153
6.2.3. Loi F (loi du rapport dispersionnel) 154
6.2.4. Remarque sur les lois non centrées du χ^2, F et t . . 157
6.2.5. Loi gamma 157
6.2.6. Loi bêta 158
6.3. Technique de simulation statistique d'observations suivant une loi donnée 160
6.3.1. Génération de nombres aléatoires uniformément répartis sur l'intervalle [0, 1] 160
6.3.2. Simulation des variables aléatoires discrètes. Méthode standard 162
6.3.3. Simulation de lois continues 163
Conclusions 165

CHAPITRE 7. Résultats fondamentaux de la théorie des probabilités . . 184

7.1. Inégalité de Tchébychev 184
7.2. Propriété de stabilité statistique des caractéristiques empiriques: loi des grands nombres et ses conséquences 185
7.2.1. Loi des grands nombres 186
7.2.2. Théorème de J. Bernoulli 187
7.2.3. Stabilité statistique des caractéristiques empiriques 187
7.3. Rôle particulier de la loi normale: théorème limite central 189
7.3.1. Théorème limite central 190
7.3.2. Théorème limite central multidimensionnel 190

7.4. Loi de probabilité de variables aléatoires qui sont fonctions de variables aléatoires connues 193
Conclusions 197

TROISIÈME PARTIE. ÉLÉMENTS DE STATISTIQUE MATHÉMATIQUE 199

CHAPITRE 8. Estimation statistique des paramètres 199
8.1. Notions préliminaires d'estimation statistique des paramètres 200
8.1.1. Position du problème 200
8.1.2. Statistiques, estimations statistiques et leurs principales propriétés 201
8.1.3. Estimation convergente 202
8.1.4. Estimation sans biais 202
8.1.5. Estimation efficace 204
8.2. Fonction de vraisemblance. Quantité d'information contenue dans n observations indépendantes par rapport à la valeur inconnue du paramètre 206
8.3. Inégalité de Rao-Cramer-Fréchet et mesure de l'efficacité des estimations 208
8.4. Propriétés asymptotiques des estimations 212
8.5. Notion d'estimation par intervalles. Construction des régions de confiance 214
8.6. Méthodes d'estimations statistiques des paramètres inconnus 215
8.6.1. Méthode du maximum de vraisemblance 215
8.6.2. Méthode des moments 223
8.6.3. Méthode des moindres carrés 226
8.6.4. Estimation par les statistiques « pondérées »; censure, troncature des échantillons et statistiques de rang comme cas particulier de pondération 231
8.6.5. Construction des intervalles de confiance (régions de confiance) 235
8.6.6. Approche bayesienne de l'estimation statistique . . . 239
Conclusions 242

CHAPITRE 9. Test d'hypothèses statistiques 245
9.1. Principaux types d'hypothèses testées 245
9.1.1. Hypothèses relatives au type de loi de probabilité de la variable aléatoire étudiée 245
9.1.2. Hypothèses d'homogénéité de deux ou plusieurs échantillons ou de certaines caractéristiques des populations analysées 246
9.1.3. Hypothèses concernant les valeurs numériques des paramètres de la population générale étudiée 247
9.1.4. Hypothèses concernant le type de dépendance entre les composantes de la variable multidimensionnelle étudiée 247
9.1.5. Hypothèses concernant l'indépendance et la stationnarité des observations traitées 248
9.2. Principe logique général d'un test statistique 248
9.3. Construction d'un test statistique; principe du rapport de vraisemblance 250
9.3.1. Essence du principe du rapport de vraisemblance . . 251
9.3.2. Test d'une hypothèse simple par le logarithme du rapport de vraisemblance 253
9.3.3. Test d'une hypothèse multiple 254

9.4. Critères de « qualité » d'un test statistique 254
9.5. Procédure séquentielle de prise de décision (tests séquentiels) 257
9.5.1. Procédure séquentielle d'observation 257
9.5.2. Test séquentiel du rapport de vraisemblance (test de Wald) et ses propriétés 258
9.5.3. Test séquentiel généralisé entre hypothèses multiples 260
Conclusions 262

QUATRIÈME PARTIE. TRAITEMENT STATISTIQUE PRIMAIRE DES DONNÉES 264

CHAPITRE 10. Statistique descriptive 264

10.1. Dossier de l'analyse; introduction et mémorisation des données; inspection des données 264
10.1.1. Dossier d'étude 264
10.1.2. Introduction et stockage des données 265
10.1.3. Inspection des données 265
10.2. Echelles de mesures 266
10.2.1. Echelle nominale 267
10.2.2. Echelle ordinale 267
10.2.3. Echelles quantitatives 268
10.2.4. Unification des données 269
10.3. Lois empiriques 271
10.3.1. Histogramme 271
10.3.2. Estimations non paramétriques de la densité . . . 273
10.3.3. Estimation de la fonction de répartition 273
10.3.4. Transformation des variables 274
10.3.5. Tableaux de contingence 275
10.4. Estimation des paramètres de localisation et d'échelle . . 276
10.4.1. Position du problème 276
10.4.2. Estimation des paramètres de la loi normale . . . 277
10.4.3. Méthode graphique d'estimation 277
10.4.4. Stabilité des estimations pour de petits écarts de la répartition par rapport à la normale 278
10.4.5. Estimation de la position du centre des lois symétriques 280
10.4.6. Paramétrisation à l'aide des estimations exponentiellement pondérées (EEP) 282
10.5. Visualisation des données multidimensionnelles 285
10.5.1. Position du problème 285
10.5.2. Composantes principales 287
10.5.3. Propriétés de la plus petite déformation de la structure géométrique des composantes principales . . . 289
10.5.4. Applications non linéaires dans un espace de moindre dimension 290
10.5.5. Analyse des proximités linéaire 293
Conclusions 294

CHAPITRE 11. Analyse préliminaire de la nature des données 297

11.1. Test d'ajustement du modèle de loi choisi avec les données initiales 297
11.1.1. Test du χ^2 de Pearson 298
11.1.2. Test de la normalité d'une répartition d'après son coefficient de dissymétrie, son coefficient d'aplatissement et ses écarts moyens 300
11.1.3. Test de Kolmogorov-Smirnov et son application à la

construction de la région de confiance pour la fonction de répartition inconnue 304
11.1.4. Test du ω^2 (Cramer-Mises-Smirnov) 306
11.1.5. Modifications des statistiques des tests de Kolmogorov-Smirnov et du ω^2 pour les échantillons de petite taille 306
11.1.6. Technique statistique de réalisation pratique des tests non paramétriques 307
11.1.7. Application des tests de Kolmogorov et du ω^2 pour vérifier la normalité d'une fonction de répartition dépendant de paramètres inconnus 308
11.2. Test des hypothèses d'homogénéité et de symétrie d'une répartition 310
11.2.1. Tests d'homogénéité basés sur les fonctions de répartition empiriques 310
11.2.2. Test d'homogénéité du χ^2 314
11.2.3. Tests de rang d'homogénéité 315
11.2.4. Test non paramétrique de l'hypothèse d'égalité des variances 319
11.2.5. Tests de rang pour $k > 2$ 321
11.2.6. Test de symétrie des répartitions 322
11.2.7. Traitement des coïncidences 324
11.2.8. Test d'homogénéité des populations normales (cas scalaire) 326
11.2.9. Test d'homogénéité de populations normales multidimensionnelles 329
11.3. Test d'indépendance et de stationnarité d'une série d'observations 332
11.3.1. Test des blocs, basé sur la médiane de l'échantillon 332
11.3.2. Test des séries « ascendantes » et « descendantes » 334
11.3.3. Test des carrés des différences séquentielles (test d'Abbe) 335
11.4. Méthodes de traitement statistique des données manquantes 336
11.4.1. Estimation des paramètres inconnus en cas d'observations manquantes 337
11.4.2. Utilisation des composantes principales 339
11.4.3. Reconstitution des données manquantes et estimation des paramètres par la méthode du maximum de vraisemblance. Estimations du « point fixe » . . . 340
11.4.4. Méthode non paramétrique d'estimation des données manquantes 343
11.5. Analyse des observations anomales 344
11.5.1. Position du problème 344
11.5.2. Méthodes graphiques 345
11.5.3. Méthode analytique d'exclusion d'une observation extrême 346
11.5.4. Test analytique d'élimination simultanée de plusieurs observations extrêmes 346
Conclusions 347

CHAPITRE 12. Logiciel de la statistique appliquée et quelques problèmes relatifs à la technique de calcul 350

12.1. Logiciel de la statistique appliquée 350
12.1.1. Organisation des paquets de programmes 351
12.1.2. Problèmes d'organisation et possibilités de gestion des données 353

12.1.3. Méthodes de traitement préliminaire (manipulation) des données 355
12.1.4. Traitement des données manquantes 355
12.1.5. Traitement primaire des données qualitatives . . 357
12.1.6. Méthodes de visualisation des données 359
12.1.7. Estimation des paramètres et localisation des observations anomales 360
12.2. Calcul des fonctions de répartition et de leurs inverses . . . 361
12.2.1. Loi normale 362
12.2.2. Loi du χ^2 364
12.2.3. Loi bêta 366
12.2.4. Loi de Fisher 367
12.2.5. Loi *t* de Student 367
12.2.6. Lois non centrées 368
12.2.7. Approximation des ailes des lois de type ω^2 . . . 369
12.2.8. Loi normale multidimensionnelle 370
12.2.9. Lois discrètes 373
12.2.10. Calcul de l'espérance mathématique des statistiques de rang 373
Conclusions 375

Notations 375

Observations primaires 375
Notions de théorie des probabilités 376
Notions de statistique mathématique 377
Bibliographie 378
Index terminologique 384

AVANT-PROPOS

La pratique des recherches en statistique appliquée montre que pour mener à bien les tâches auxquelles il est confronté le chercheur doit se sentir à l'aise dans trois domaines:

l'appareil mathématique de la statistique appliquée, c'est-à-dire essentiellement les méthodes appliquées de statistique mathématique;

la modélisation mathématique, c'est-à-dire l'art de formaliser le problème posé, un art qui consiste à traduire le problème du langage de la discipline utilisatrice (économie, sociologie, médecine, etc.) dans le langage des schémas et modèles mathématiques abstraits;

le logiciel respectif.

Les monographies, manuels et guides de traitement statistique des données publiés en U.R.S.S. et dans les autres pays n'étudient en principe que l'un des trois domaines mentionnés *). Cette situation s'explique à notre sens dans une grande partie par le fait que le point faible des applications statistiques réside dans la distance significative qui sépare l'instant de la mise au point de l'appareil mathématique de l'instant où cet appareil commence à être utilisé avec efficacité dans la résolution des problèmes posés.

Le principal objectif des auteurs de cet ouvrage est de proposer un exposé systématique et cohérent sur les trois aspects (énumérés plus haut) de l'analyse statistique de l'information initiale au niveau de la première étape de traitement des données empiriques. Cette étape, appelée dans cet ouvrage étape de traitement statistique primaire, comprend: le dossier d'étude; l'organisation de l'en-

---

*) Ces trois à cinq dernières années, de nombreux ouvrages étrangers ont été publiés qui décrivent les méthodes de traitement statistique en même temps que le logiciel correspondant. Voir par exemple E. Diday et collaborateurs. *Optimisation en classification automatique*. Paris, INRIA, 1980; Lee Tc., Judge G., Zellner A. *Estimating the parameters of the Markov probability model from aggregate time series data*. North-Holland Publishing Company, Amsterdam-London, 1970.

trée et la mémorisation des données dans l'ordinateur; la visualisation des données multidimensionnelles en vue de la formulation des hypothèses de travail; la manipulation des variables de divers types; le traitement des données manquantes et des données anomales; la description et l'analyse des répartitions empiriques; le test des hypothèses d'indépendance, d'homogénéité et de stationnarité des échantillons.

L'ouvrage se compose de quatre parties. Les principes méthodologiques de la statistique appliquée et les principales définitions et méthodes de la modélisation font l'objet de la première partie. Cette partie est la moins « mathématisée », mais son assimilation permettra au lecteur de bien maîtriser les méthodes de traitement statistique des données. Les parties II et III (chapitres 4 à 9) développent les notions fondamentales de théorie des probabilités et de statistique mathématique. Ces chapitres peuvent servir aux étudiants des facultés scientifiques. L'assimilation de ce matériel exige du lecteur la connaissance des seules notions élémentaires de mathématiques, notamment les notions de fonction et de graphique d'une fonction; de suite numérique et de somme d'une suite; de dérivée (ordinaire et partielle); d'intégrale (définie et indéfinie); de calcul vectoriel et matriciel.

La partie IV enfin est consacrée à la description des méthodes de traitement statistique préliminaire des données (chap. 10 et 11), aux aspects numérique et informatique de leurs réalisations sur ordinateur (chap. 12).

Les tâches ont été réparties comme suit entre les auteurs: Aïvazian: Avant-propos, chap. 1, 2, 4 à 9, §§ 3.2, 11.3, n°s 11.1.1, 11.1.2 et 11.2.8; Enukov: chap. 11, 12, § 10.5 (en collaboration avec Méchalkine), n° 8.6.6; Méchalkine: chap. 3, 10, §§ 6.3, 11.5, ainsi que (en collaboration avec Aïvazian) n°s 4.2.1, 5.4.3, 6.1.11, 6.2.4, § 8.4, n°s 8.6.3, 9.3.2 et 9.3.3.

La nécessité de la publication d'un tel ouvrage tient à nos yeux aux raisons objectives suivantes.

Primo, ces quinze à vingt dernières années, on a assisté à un intense développement des méthodes de traitement statistique qui font une large place à l'informatique tout en affaiblissant considérablement les hypothèses concernant le modèle mathématique et allant même jusqu'à renoncer à toute conjecture *a priori* sur la nature probabiliste des données traitées (statistique non paramétrique, modèles non linéaires et estimation itérative des paramètres, procédures statistiques robustes, méthodes non probabilistes d'analyse des données).

Secundo, le logiciel de la statistique mathématique appliquée est assez bien élaboré en U.R.S.S. comme dans d'autres pays: signalons entre autres le PPSA [108], l'OTEKS [106], les bulletins (vypousky) de logiciel du système unifié des ordinateurs de l'Institut de mathématiques de l'Académie des sciences de Biélorussie [88], le

système de programmes de reconstitution des dépendances au vu des échantillons de taille limitée de l'Institut des problèmes de gestion de l'Académie des sciences de l'U.R.S.S., le système de traitement des données empiriques SITO de l'Institut de recherches scientifiques de Léningrad du Centre de calcul de l'Académie des sciences de l'U.R.S.S. et les paquets étrangers *BMDP*, *SAS*, *SPSS*, *STIL*, *OSIRIS*, *TROLL*, le logiciel de traitement statistique multidimensionnel (*CNET*, Paris, 1982), *SPAD*-système portable pour l'analyse des données (*ŒSIA*, Paris, 1982) et autres *). Mais les ouvrages actuels de statistique mathématique et traitement statistique des données (parmi lesquels il faut particulièrement signaler l'ouvrage en trois volumes de Kendall M. et Stuart A. [76], [77], [78]) escamotent les méthodes non paramétriques et, rappelons-le, le logiciel de statistique mathématique appliquée. Les recueils de programmes statistiques (cf. [24], [88], [106], [108], [116]) ne contiennent aucun exposé tant soit peu systématique de ces méthodes.

Dans cet ouvrage on a inclus de nombreux résultats inédits: la formulation d'optimisation des problèmes de statistique appliquée (§ 1.2), la comparaison des bonnes propriétés des tests de Wald et de Neyman-Pearson (§ 9.5), la méthode des moments de pondération exponentielle pour l'établissement d'estimations statistiques robustes (§ 10.4), quelques méthodes de « codage » des variables qualitatives (§§ 10.2, 10.5, n° 12.1.5).

Cet ouvrage est le fruit des activités scientifiques et pédagogiques déployées par les auteurs à l'Institut central d'économie et de mathématiques de l'Académie des sciences de l'U.R.S.S., à l'Université d'Etat de Moscou et au Département central du Ministère de la Santé de l'U.R.S.S.

Les auteurs expriment leur gratitude à V. Fedorov et E. Yassine qui se sont chargés de la lecture du manuscrit et ont contribué à améliorer l'exposé par leurs remarques judicieuses. Le contenu et la logique de cet ouvrage ont fortement bénéficié des contacts permanents des auteurs avec leurs collègues du séminaire scientifique « Analyse statistique multidimensionnelle et modélisation probabiliste des processus réels » organisé sous les auspices du Conseil scientifique de l'Académie des sciences de l'U.R.S.S. sur le problème « Planification optimale et gestion de l'économie » depuis 1969, ainsi que du séminaire méthodologique national « Problèmes numériques de statistique mathématique » qui se tient à l'Université de Moscou sous la direction de Yu. Prokhorov.

Les auteurs sont reconnaissants à L. Mett pour la présentation du manuscrit.

Les auteurs

*) Pour plus de détails sur ces travaux, consulter l'ouvrage: *Algorithmes et logiciel d'analyse statistique appliquée*. Outcheni zapiski po statistiké, t. 36, M., Naouka, 1980 (en russe).

PREMIÈRE PARTIE

STATISTIQUE APPLIQUÉE : ESSENCE ET OBJECTIF

(principes méthodiques généraux)

CHAPITRE PREMIER

LA STATISTIQUE APPLIQUÉE : UNE DISCIPLINE SCIENTIFIQUE AUTONOME

1.1. Lien entre la statistique appliquée et les autres disciplines statistiques et principales étapes de la recherche statistique

1.1.1. Définition de la statistique appliquée. Faut-il user de ce terme ou se contenter de la traditionnelle notion de « statistique mathématique »? Comment la statistique appliquée est-elle reliée aux autres disciplines statistiques telles la « statistique mathématique », l'« analyse des données », la « statistique économique », etc. ? Deux circonstances au moins plaident en faveur de la légitimité et de l'adéquation de traiter la statistique appliquée comme une discipline scientifique autonome.

Primo, jusqu'alors deux tendances parallèles se sont dégagées dans le développement de la théorie, de la méthodologie et de la pratique du traitement statistique des données analysées. La première met en jeu des méthodes *qui envisagent la possibilité d'une interprétation probabiliste* des données traitées et des résultats statistiques fournis par le traitement. Ces méthodes (et elles seules !) font la matière de l'écrasante majorité des monographies et traités de statistique mathématique. En d'autres termes, par méthodes de statistique mathématique on comprend les seules méthodes de traitement statistique des données initiales, méthodes qui sont élaborées et mises en œuvre sous le rapport de la nature probabiliste de ces données *). La deuxième tendance fait intervenir une classe assez vaste de méthodes de traitement statistique de l'information initiale, plus exactement, tout l'ensemble des méthodes qui *a priori ne s'appuient pas sur la nature probabiliste des données traitées* (telles sont les méthodes *de classification automatique*, *d'analyse des proximités*, *de la théorie des*

*) Telle est la situation *de facto*. Mais *de jure* si l'on se refère à l'Encyclopédie soviétique, la statistique mathématique est comprise dans un sens plus large, plus exactement comme une « science sur les méthodes mathématiques de systématisation et d'utilisation des données statistiques pour des conclusions scientifiques et pratiques. De nombreuses sections de la statistique mathématique se basent sur la théorie des probabilités qui permet d'estimer la fiabilité et l'exactitude des conclusions tirées à partir des données statistiques limitées ».

mesures, etc.) ne s'inscrit pas dans le cadre de la statistique mathématique.

Secundo, les spécialistes de l'élaboration et de l'application des méthodes de traitement statistique de l'information initiale ne peuvent tirer un trait sur l'importante étape qui sépare la fin de l'élaboration d'une méthode purement mathématique de l'instant où cette méthode commence à donner ses fruits. Au cours de cette étape le mathématicien appliqué devra :

bien appréhender l'essence du problème, utiliser de façon adéquate les hypothèses initiales (sur lesquelles repose toute méthode mathématique) pour comprendre la nature du problème réel ;

résoudre (dans certains cas spéciaux) un problème fort complexe de réduction de l'information initiale présentée, par exemple, sous forme de signaux physiques, de balayages radar, de coupes géologiques, etc. à une forme standard (unifiée) (cf. (1.4) et (1.4′)) ;

composer des algorithmes *réalisables* et le logiciel en tenant compte de la nature de l'information statistique traitée et des performances des ordinateurs ;

mettre sur pied un procédé commode et efficace de communication avec l'ordinateur durant la résolution du problème.

Les notions, les méthodes et les résultats qui permettent de franchir cette étape, ainsi que l'ajustement et la mise au point de l'outillage mathématique, constituent l'essentiel de la statistique appliquée. Ainsi donc, la *statistique appliquée est une discipline autonome élaborant et systématisant des notions, des recettes, des méthodes et des modèles mathématiques destinés à l'organisation de la collecte* *), *à l'écriture standard, à la systématisation et au traitement (par ordinateur) des données statistiques afin de les représenter sous une forme commode, de les interpréter et d'en tirer des conclusions scientifiques et pratiques.*

Certains spécialistes se servent pour définir ce système de notions, recettes, méthodes et modèles mathématiques du terme « analyse des données » pris dans une acception large.

1.1.2. Deux façons d'interpréter les données initiales et deux approches de leur traitement statistique. Considérons deux exemples. Dans le premier, le but de l'analyse statistique est d'étudier les capacités d'une production en série sur des données initiales qui

*) Par « organisation de la collecte » des données statistiques on comprend uniquement la détermination du procédé de sélection des unités (familles, entreprises, pays, malades, etc.) à étudier statistiquement dans une population (cf. description de l'étape 2 dans 1.1.3). Nous n'incluons pas ici l'élaboration d'une méthodologie et d'un système de variables de l'objet étudié : ce travail implique une approche *professionnelle* (économique, technique, médicale, etc.) de problèmes dont la résolution exige une information statistique, une approche qui est donc du ressort des spécialistes compétents et qui peut constituer un domaine spécial de recherches (par exemple, un domaine de la statistique économique).

sont le résultat du contrôle (d'après le principe binaire) d'un certain nombre d'articles prélevés aléatoirement. Si l'on teste n articles, les résultats peuvent être représentés sous la forme d'une suite de nombres

$$x_1, x_2, \ldots, x_n, \tag{1.1}$$

où le résultat du contrôle de l'article x_i est supposé égal à 1 si cet article est défectueux et à 0 dans le cas contraire. Si la production est au point et fonctionne en régime stationnaire (c'est-à-dire que ses capacités technologiques restent à un niveau constant), il est alors naturel d'interpréter la suite d'observations (1.1) comme un échantillon fini de la population infinie que nous aurions obtenue si l'on avait contrôlé tous les articles d'affilée. Dans ce cas, cet échantillon est considéré comme une partie intégrante, comme un *représentant* d'une population infinie « située après lui », et ses principales caractéristiques statistiques, par exemple sa moyenne arithmétique, qui on le voit aisément est la partie d'articles défectueux de cet échantillon, comme une approximation du nombre exact d'articles défectueux de toute la production. En pareille circonstance, on a la possibilité virtuelle de répéter l'observation (ou l'expérience) réalisée en gardant le même ensemble réel de conditions, ensemble qui fait la part de l'action « perturbatrice » d'un grand nombre de facteurs aléatoires incontrôlables (qui sont la cause de la *stochasticité* du résultat de chaque observation). De telles situations peuvent être décrites par un *modèle probabiliste* (cf. § 1.2 et 1.3). La suite d'observations (1.1) s'interprète alors comme un échantillon aléatoire d'une certaine population, c'est-à-dire comme les valeurs expérimentales (ou observées) de la variable aléatoire étudiée, et son traitement statistique implique des méthodes statistiques classiques (méthodes d'estimation statistique des paramètres inconnus, méthodes de test des hypothèses statistiques, etc., cf. troisième partie).

Dans le cadre de cette interprétation probabiliste des données statistiques initiales (ou *fichier*), le chercheur est confronté simultanément à deux populations d'articles : une *population réellement observée* représentée statistiquement par une suite d'observations (1.1) (c'est-à-dire un *échantillon*) et une population théoriquement imaginable (dite *population générale*). Les principales propriétés et caractéristiques de l'échantillon dites empiriques peuvent être analysées et calculées à partir des données statistiques (1.1). Les principales propriétés et caractéristiques de la population générale, dites théoriques, ne sont pas connues du chercheur, et la raison d'être des méthodes de statistique mathématique est justement de permettre de les déterminer avec le plus de précision à partir de celles des échantillons.

Illustrons la deuxième interprétation des données statistiques initiales sur l'exemple suivant. Une étude statistique a été réalisée sur un ensemble de 74 villes moyennes (de 100 000 à 500 000 habitants) de la Fédération de Russie. Pour chaque ville on a relevé les

valeurs de 32 variables $x^{(1)}, x^{(2)}, \ldots, x^{(32)}$, caractérisant le niveau d'instruction, la composition d'après le milieu social, l'âge et le sexe, la structure de l'emploi des habitants *). Les données statistiques initiales peuvent être représentées ici par une suite de vecteurs à 32 dimensions

$$X_1, X_2, \ldots, X_{74}, \tag{1.2}$$

où le résultat du sondage de la ville i est le vecteur

$$X_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(32)} \end{pmatrix}, \tag{1.3}$$

dont les composantes sont les valeurs numériques des variables analysées pour chaque ville. Le but de l'analyse statistique des données initiales (1.2) est de déterminer le nombre et la composition des divers types de villes, sous-entendu que par type on comprend une classe de villes de l'ensemble étudié, homogènes (similaires) par la structure du niveau d'instruction, de la composition par âge et par sexe et du caractère socio-professionnel de leurs habitants**). Si l'on admet que la proximité géométrique de deux points-villes X_i et X_j de la forme (1.3) dans l'espace respectif à 32 dimensions traduit leur homogénéité (similitude) par rapport aux variables analysées et justifie leur appartenance à un même type, alors pour résoudre le problème posé ci-dessus il faut faire intervenir des méthodes adéquates de classification automatique (reconnaissance des images « sans maître ») et d'abaissement de l'ordre. L'appareil mathématique de ces méthodes prévoit certes le calcul de caractéristiques telles les moyennes,

*) Chaque variable $x^{(k)}$ ($k = 1, 2, \ldots, 32$) représente le nombre moyen d'habitants pour mille doués d'une certaine caractéristique : les quatre premières variables ($x^{(1)}$ à $x^{(4)}$) représentent le niveau d'instruction (supérieure, supérieure non achevée, secondaire spéciale ou brevet) ; les douze suivantes ($x^{(5)}$ à $x^{(16)}$), la composition en fonction de l'âge et du sexe ; les cinq suivantes ($x^{(17)}$ à $x^{(21)}$), le caractère socio-professionnel de la population ; les autres, enfin, la répartition de la population suivant les branches de production de biens matériels et non matériels et suivant les sources de revenu.

**) Ce problème s'est posé comme une étape intermédiaire d'une analyse assez détaillée des données statistiques sur les villes russes, visant à dégager les traits les plus caractéristiques et les lois de la structure socio-économique d'une ville moyenne type. La réalisation d'une analyse circonstanciée dans chacune des 74 villes (sondages et enquêtes auprès de la population) est une œuvre de Titan. Aussi a-t-on opté pour une solution qui impliquait de subdiviser l'ensemble des villes étudiées en un certain nombre de types (de classes homogènes) dans l'espace à 32 dimensions analysé ; on a repéré dans chaque classe les villes les plus caractéristiques : les villes étalons (par exemple les villes les plus proches du « centre de gravité » de leurs classes), puis on les a sélectionnées pour un examen plus détaillé. Cet examen a préludé aux recherches socio-économiques « Taganrog-1 », « Taganrog-2 ».

les variances, les covariances, etc., mais celles-ci caractériseront la nature et la structure seulement des données réellement analysables, c'est-à-dire seulement de l'ensemble des 74 villes analysées statistiquement. Contrairement à l'exemple précédent qui impliquait l'analyse statistique des résultats du contrôle d'articles produits en régime stationnaire, l'exemple envisagé nous confronte à de sérieuses difficultés méthodologiques au niveau de

l'interprétation des données statistiques initiales (1.2) comme un échantillon d'une population générale (théoriquement imaginable);

l'utilisation des modèles probabilistes pour l'élaboration et le choix des meilleures méthodes de traitement statistique;

l'interprétation probabiliste des conclusions déduites de l'analyse statistique des données initiales.

C'est là que réside la différence fondamentale entre les deux approches de l'analyse statistique des données initiales. Mais dans l'une comme dans l'autre, le choix de la meilleure méthode de traitement s'effectue selon un critère matérialisé par une *fonctionnelle de qualité de la méthode*. Les approches décrites se différencient par le *procédé de justification* du choix de la fonctionnelle de qualité, ainsi que par l'interprétation de cette fonctionnelle et des conclusions statistiques établies: dans le premier cas, l'analyste fonde son choix sur les hypothèses relatives à la nature probabiliste des données initiales et utilise lesdites hypothèses pour l'interprétation probabiliste de ses conclusions; dans le second cas, l'analyste ne dispose d'aucune information *a priori* sur la nature probabiliste des données initiales et pour justifier son choix de la fonctionnelle de qualité il en appelle à des considérations d'ordre physique: comment et à quelle fin ont été obtenues les données à traiter. Mais une fois choisi le critère d'optimisation de la qualité de la méthode, les outils mathématiques de résolution du problème de statistique envisagé sont communs aux deux approches: dans les deux cas, l'analyste utilise des méthodes de résolution de problèmes d'extremums. A vrai dire, ces approches divergent de nouveau au stade final de l'interprétation des résultats obtenus.

Donc, ces deux approches ont en commun l'information statistique initiale et la recherche du meilleur (sur le plan de l'optimisation d'une certaine fonctionnelle de qualité de la méthode) traitement statistique de cette information pour en extraire des résultats scientifiques ou pratiques.

Ainsi, avant d'entamer une analyse statistique de données initiales l'analyste doit tout d'abord définir dans le cadre de quelle approche il entend procéder à son analyse. En d'autres termes, il doit se livrer à un *choix fondamental du type de modèle*. A cet égard, les mises en garde de certains auteurs (cf. [12], [134]) contre la nocivité d'utiliser en toute occasion (parfois de façon irréfléchie) les méthodes probabilisto-statistiques comme principal instrument de traitement

des données initiales nous semblent judicieuses et utiles. Mais ceci ne doit pas jeter le discrédit sur ces méthodes : or, c'est semble-t-il l'objectif que s'est fixé l'auteur de [12] et c'est à cette conclusion (c'est-à-dire à l'inutilité et l'inefficacité des méthodes statistiques probabilistes) que sont arrivés de nombreux lecteurs de [134] que l'auteur l'ait voulu ou non.

En réalité, il nous faut partir de la situation suivante. Nous commencerons à partir de l'instant où l'analyste est en possession des données statistiques initiales caractérisant les divers aspects du processus ou de l'événement étudié. Le problème est de traiter cette information de la meilleure façon possible (dans un certain sens) pour en tirer des conclusions scientifiques ou pratiques d'une certaine nature sur l'événement étudié. Pour préciser ce qu'il entend par « meilleure façon possible », l'analyste doit formaliser son problème, *choisir un modèle*. Tout modèle est une représentation (mathématique) simplifiée de la réalité étudiée (cf. § 3.1). Il est évident que l'adéquation du modèle choisi et de la réalité étudiée est un facteur décisif garant de l'efficacité des méthodes de traitement statistique employées. Comme aucun des modèles sélectionnés ne reflète parfaitement la réalité, on ne peut qu'applaudir à la propension de l'analyste à reprendre le traitement de ses données initiales en modifiant légèrement le modèle à chaque fois (cf. développement de cette démarche au § 1.2).

1.1.3. Principales étapes du traitement statistique des données initiales. Essayons maintenant de décrire le schéma logique général d'analyse statistique des données initiales. Pour expliquer le rôle et la place des principales recettes de modélisation statistique et des méthodes de traitement primaire des données initiales, nous décomposerons ce schéma en étapes pour la commodité. Il va de soi que cette décomposition est purement conventionnelle. En particulier, elle ne signifie pas que les étapes doivent respecter obligatoirement l'ordre chronologique. Bien plus, de nombreuses étapes (par exemple les étapes 4, 5 et 6) se trouvent en rapport d'itération, c'est-à-dire que les résultats de la réalisation d'étapes ultérieures peuvent contenir des conclusions impliquant la répétition (compte tenu de la nouvelle information) des étapes précédentes.

Etape 1 : analyse initiale (préliminaire) du système étudié. Cette analyse définit : a) les principaux objectifs de l'étude à un niveau non formalisé dans les termes de la discipline utilisatrice ; b) l'ensemble des unités faisant l'objet de l'étude statistique ; c) la liste $(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ des variables qui caractérisent l'état (le comportement) de chaque objet sondé (ces variables sont sélectionnées dans un ensemble préparé par les spécialistes de la discipline utilisatrice) ; d) le degré de formalisation des notations pendant la collecte des données ; e) la durée totale et le travail nécessités pour les

opérations projetées, la durée et le volume du sondage; f) les situations impliquant une vérification préalable avant la composition d'un plan d'étude détaillé (par exemple, l'identification des unités d'observation n'est pas toujours évidente *a priori*, ou dans un autre ordre d'idées l'accord d'un malade de se conformer aux prescriptions du médecin n'est pas automatiquement acquis, etc.); g) la position formalisée du problème avec éventuellement le modèle probabiliste de l'événement étudié et la nature des conclusions statistiques auxquelles doit (ou peut) aboutir l'analyste par traitement des données initiales; h) les formes de collecte de l'information primaire et de son introduction dans l'ordinateur.

La première étape nécessite d'une équipe compétente un effort parfois aussi important que pour toutes les autres étapes, sous réserve que le traitement soit effectué avec un paquet de programmes convenables *). Ceci désigne tout particulièrement les méthodes *d'assistance de l'ordinateur* pour la réalisation de cette partie. Cette assistance peut consister en une suggestion (accompagnée d'une estimation) des formes de documentation pour la collecte de l'information initiale, des méthodes de construction des groupes de contrôle ou de « pseudo-contrôle » lors de l'étude d'une action quelconque (ce qui est d'actualité pour les applications en médecine), des modèles convenables, du dictionnaire, etc.

Etape 2: *composition d'un plan circonstancié de collecte de l'information statistique initiale.* La composition de ce plan doit dans la mesure du possible tenir nécessairement compte du schéma tout entier de l'analyse statistique, ce qu'on oublie souvent. La collecte de l'information gagnerait beaucoup si l'on savait *a priori* comment et à quelle fin se fait l'analyse. Une attention toute particulière doit être portée aux cas où: a) on se sert de l'appareil de l'analyse des échantillons (cf. par exemple [36]), c'est-à-dire qu'on définit la nature de l'échantillon (aléatoire, proportionnel, stratifié, etc.); b) on calcule le « pouvoir de résolution » d'une étude d'un volume et d'une durée donnés (cf. par exemple [95] où on majore le nombre d'associations statistiquement significatives entre les facteurs risques et la fréquence des maladies ou bien [35] où on propose un modèle élémentaire pour la description phénoménologique de l'action des soins); c) l'expérience est active pour une partie au moins des variables d'entrée : dans chaque observation concrète les variables sont fixées à un certain niveau et le plan d'étude est défini à l'aide des méthodes de planification des expériences (de régression) (cf. par exemple [52]). Dans certains traités de théorie générale de statistique (cf. par exemple [131]) cette étape est dite étape de « préparation organisationnelle et méthodique ». Comme déjà signalé, les problèmes d'élaboration

*) Dans certaines études sociologiques spéciales, médicales et statistiques caractérisées par une grande dépense en temps et en moyens pour la collecte de l'information, la thèse formulée n'est valable que si l'on exclut l'étape 3.

d'une méthode de détermination d'un système de variables initiales caractérisant l'objet ou le processus étudié n'ont pas été inclus dans les étapes décrites ici et relèvent de la discipline utilisatrice du sondage (économie, médecine, etc.).

Etape 3 : collecte des données statistiques initiales et leur introduction dans l'ordinateur. On introduit en même temps les définitions complètes ou sommaires (à des fins de reproduction automatique sur des tables) des termes employés. Le paquet doit prévoir des mesures excluant ou minimisant le risque de calcul sur un sous-ensemble inadéquat de données ou pour un autre sous-groupe d'objets.

Donc, que l'analyste choisisse en personne la méthode et le plan d'étude statistique ou qu'il soit en possession des résultats de *l'expérience passive*, au moment où il aura à définir les principaux instruments de l'étude statistique, il disposera en général des données initiales sous la forme d'une suite, dépendant du temps, de matrices d'observations

$$\mathbf{X}(t) = \begin{Bmatrix} x_1^{(1)}(t), & x_2^{(1)}(t), & \ldots, & x_n^{(1)}(t) \\ x_1^{(2)}(t), & x_2^{(2)}(t), & \ldots, & x_n^{(2)}(t) \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(p)}(t), & x_2^{(p)}(t), & \ldots, & x_n^{(p)}(t) \end{Bmatrix} \quad (t = t_1, \ldots, t_N), \qquad (1.4)$$

où $x_i^{(k)}(t)$ est la valeur de la variable k qui caractérise l'état de l'objet i à l'instant t. Il arrive que les t_l soient aléatoires pour chaque objet. Cette situation se présente par exemple *en médecine* lorsque le vecteur

$$X_i(t_l) = \begin{pmatrix} x_i^{(1)}(t_l) \\ \vdots \\ x_i^{(p)}(t_l) \end{pmatrix}$$

caractérise le cours de la crise l du malade i et les malades ne présentent pas le même nombre de crises durant le même intervalle de temps $[0, T]$. Dans ce cas les matrices $(X_i(t_l))_{t_l \in [0, T]}$ ont des dimensions différentes pour des malades différents (c'est-à-dire pour des i différents). En médecine certaines coordonnées sont représentées non pas par des chiffres mais sous forme d'un texte. Ce mode de représentation des données initiales est propre aux études *sociologiques*, et dans un degré moindre aux études *économiques*.

Souvent, et plus particulièrement lorsque les données statistiques initiales résultent de sondages, d'enquêtes, d'expertises, il est possible que l'observation primaire porte non pas sur l'état de l'objet i à l'instant t, mais sur la caractéristique $\rho_{ij}(t)$ de la proximité (l'éloignement) de deux objets (ou de variables) de numéros respectifs i et j à l'instant t. Dans ce cas les données initiales sont représentées par une suite dépendant du temps de matrices d'ordre n (si l'on considère les caractéristiques de proximité de deux *objets*) et d'ordre p

(si l'on considère les caractéristiques de proximité de deux *variables*) de la forme

$$\rho(t) = \begin{pmatrix} \rho_{11}(t), & \rho_{12}(t), & \ldots, & \rho_{1m}(t) \\ \rho_{21}(t), & \rho_{22}(t), & \ldots, & \rho_{2m}(t) \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{m1}(t), & \rho_{m2}(t), & \ldots, & \rho_{mm}(t) \end{pmatrix} \quad \begin{pmatrix} m = n \text{ ou } p; \\ t = t_1, \ldots, t_N \end{pmatrix}. \qquad (1.4')$$

Il est évident qu'on peut passer directement de (1.4) à (1.4') (sous réserve que soit définie une métrique sur l'espace des objets et sur celui des variables). En général il est impossible de passer inversement de (1.4') à (1.4) sans conditions supplémentaires et sans méthodes spéciales (par exemple celles de l'analyse des proximités [84]). Il existe d'autres formes de représentation de la structure géométrique des données initiales, mais nous les négligerons ici.

Pour simplifier les notations, on considérera, sauf mention expresse du contraire, la variante *statique* du schéma, c'est-à-dire une situation dans laquelle on considérera les données initiales (1.4) ou (1.4') *à un instant fixe t* que l'on omettra dans les notations.

Etape 4 : traitement statistique primaire des données. Le traitement statistique primaire des données doit généralement résoudre les problèmes suivants : a) transformation des variables textuelles en variables nominales (avec un nombre assigné de modalités) ou ordinales ; b) description statistique des populations de base avec indication des limites de variation des variables ; c) analyse des observations anomales ; d) reconstitution des données manquantes ; e) vérification de l'indépendance statistique des données initiales ; f) unification des types de variables par des procédés divers ; g) analyse expérimentale de la loi de probabilité de la population générale étudiée et paramétrisation de l'information recueillie sur la nature des répartitions étudiées (cette étape est parfois appelée *établissement de la liste des indices agrégés* [131]). De plus, dans le cadre de l'étape 4, il faut définir la dimension du problème, la complexité de son algorithme, l'aptitude de l'ordinateur à le résoudre, sa formulation dans le langage d'entrée du paquet, etc. (pour plus de détails voir la description de l'étape 6).

Appesantissons-nous sur certaines des questions abordées.

Analyse des observations anomales. Parfois un examen (automatique ou visuel) même fugitif des données initiales (1.4) ou (1.4') peut jeter la suspicion sur la véracité (ou la légitimité) de certaines observations qui tranchent nettement sur les autres. Dans ces cas, il est légitime de se poser la question suivante : ces écarts considérables (dépassements anomaux) sont-ils le fait des fluctuations aléatoires habituelles de l'échantillon (fluctuations conditionnées par la nature de la population générale analysée) ou bien le fait d'une transgression aux normes habituelles de recueil des données statistiques, voire d'une erreur flagrante d'enregistrement (d'écriture) ? Dans les

deux derniers cas les observations « suspectes » doivent être manifestement bannies.

Le seul moyen absolument sûr de supprimer les observations anomales est d'étudier en détail les conditions dans lesquelles elles ont été obtenues. Mais dans bien des cas cette analyse est objectivement difficile ou pratiquement impossible. Force est donc de se rabattre sur des méthodes (statistiques) formelles adéquates. Le schéma logique général de ces méthodes est le suivant : en partant d'une hypothèse de base sur la nature de l'ensemble des données analysées, l'analyste se donne une fonction

$$\psi(X^*_{i_1}, X^*_{i_2}, \ldots, X^*_{i_k}\,;\, \mathrm{X}) \tag{1.5}$$

de toutes les observations X, caractérisant le degré d'anomalie (l'écart par rapport à la principale masse des observations « suspectes » $X^*_{i_1}, \ldots, X^*_{i_k}$, porte dans (1.5) les valeurs réelles des observations et compare la valeur obtenue à un seuil ψ_0 *); si $\psi > \psi_0$, les valeurs suspectes soit sont définitivement exclues, soit leur contribution est diminuée par le biais d'une fonction de poids qui décroît lorsque le degré d'anomalie des observations croît.

Diverses méthodes d'analyse des observations anomales sont accessibles dans le § 11.5 (cf. également [7], [124]).

Reconstitution des observations manquantes (effacées). Dans les matrices des données statistiques initiales (1.4) ou (1.4') on constate parfois pour des raisons diverses (y compris après l'exclusion des observations anomales) des *omissions* d'éléments isolés ou de toute une partie des lignes ou des colonnes. L'élimination *ipso facto* de l'objet tout entier (de la colonne incomplète) ou de l'indice (la ligne incomplète) nous priverait d'une information précieuse. Il se pose donc la question d'une meilleure reconstitution des données manquantes (effacées). Le critère de qualité de la reconstitution des données manquantes dépend du caractère du traitement ultérieur des données initiales, c'est-à-dire de la finalité de l'étude (cf. § 11.4 ainsi que [19], [106], [142]).

Vérification de l'homogénéité de quelques tranches de données initiales. Les conditions objectives de recueil des données statistiques notamment dans le cas d'une expérience passive peuvent être telles que la $(p \times n)$-matrice des observations (1.4) est la juxtaposition de $(p \times n_1)$-, $(p \times n_2)$-, ..., $(p \times n_k)$-matrices respectives $\mathrm{X}_1, \ldots, \mathrm{X}_k$ d'observations partielles $(n_1 + n_2 + \ldots + n_k = n)$, où chaque matrice partielle X_j définit une tranche de données initiales se rapportant à une sous-population de n_j objets. Ceci étant, le

*) Dans la position probabiliste du problème, le seuil ψ_0 est défini par des tables statistiques en fonction de la loi de probabilité de la statistique ψ sous réserve que les observations $X^*_{i_1}, \ldots, X^*_{i_k}$ soient « blanchies ».

sondage (sa date) de ces populations peut être échelonné dans l'espace (dans le temps).

Il est évident qu'avant de traiter statistiquement les données initiales (c'est-à-dire avant de leur appliquer telle ou telle méthode d'analyse en fonction de l'objectif final de l'analyse) l'analyste doit répondre à une question : la réunion des tranches (échantillons) en un seul bloc est-elle légitime ou bien chaque tranche est-elle spécifique et doit-elle par conséquent être traitée séparément ? Au niveau des modèles statistiques mathématiques cette question consiste à établir (par des tests statistiques adéquats) si l'on peut traiter les tranches de données $\mathbf{X}_1, \ldots, \mathbf{X}_k$ comme des *échantillons parents* (i.e. extraits d'une même population générale ; cf. par exemple [16]). Il est évident que si la cause du fractionnement est l'échelonnement des tranches respectives dans le temps, alors on a affaire à une vérification de la stationnarité des séries temporelles multidimensionnelles correspondantes.

Vérification de l'indépendance statistique d'une suite d'observations constituant les données initiales. De nombreuses méthodes statistiques ne s'appliquent qu'à des suites d'observations $X_1, X_2, \ldots, X_n$ indépendantes. Ceci concerne également les suites $\{X_i(t_1), \ldots, X_i(t_N)\}$. Donc, avant de soumettre les résultats des observations au traitement statistique principal, il faut déterminer (par des tests statistiques adéquats (cf. § 11.3)) s'ils sont statistiquement indépendants ou s'il faut les traiter comme des suites de quantités interdépendantes.

Unification des variables. L'une des difficultés de l'analyse automatique de l'information réside dans le fait que les composantes $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ de la variable multidimensionnelle analysée peuvent être de trois types : *quantitatives*, *qualitatives* (ordinales) et *nominales*. Leur définition et les principales formes d'écritures de leurs valeurs observées sont citées dans les §§ 5.3 et 10.2.

Ceci pose la question de l'uniformisation de l'enregistrement d'une observation de l'objet i. Une solution consiste à représenter la i-ième observation multidimensionnelle par un vecteur colonne de dimension $m_1 + m_2 + \ldots + m_p$, où m_k est le nombre de modalités (d'intervalles de groupement, de niveaux de qualité ou de groupes homogènes) de la variable $x^{(k)}$, dont les composantes ne sont que des 0 et des 1. Si l'on opte pour ce procédé d'uniformisation, on doit prendre notre parti *primo* du subjectivisme qui préside au choix de la méthode de subdivision des intervalles de variation des variables analysées en sous-intervalles de groupement, *secundo* d'une certaine perte d'information sur les données initiales, entraînée par le passage des valeurs *individuelles* aux valeurs *groupées* des variables quantitatives.

On pourrait utiliser une autre méthode d'uniformisation de l'enregistrement des données initiales dont le principe est diamétrale-

ment opposé au précédent. Plus exactement, en s'appuyant sur des considérations (et des hypothèses) complémentaires, l'analyste peut tenter de transformer les variables qualitatives et nominales en variables quantitatives par « codage » ou par une méthode de l'analyse des proximités, ainsi que par des modèles spéciaux (de Terstown, Lazarsfeld et autres), cf. § 10.2 et [48], [98], [118], [121].

Analyse expérimentale et paramétrisation de la loi de probabilité de la population générale étudiée. Cette partie du traitement statistique préliminaire des données initiales (1.4) comprend le calcul des principales caractéristiques numériques de la répartition : la valeur moyenne, la variance, les coefficients de dissymétrie et d'aplatissement, et dans le cas multidimensionnel, les éléments de la matrice des covariances empiriques. Par ailleurs, l'analyste procède à une analyse numérique et graphique des lois de probabilité à une dimension des variables étudiées en construisant les polygones de fréquence, les histogrammes et les fonctions de répartition empiriques respectifs. Les résultats de cette analyse expérimentale complétés par les informations dont on dispose sur la nature de la population générale analysée peuvent parfois suffire pour énoncer une ou plusieurs hypothèses concurrentes sur la forme (paramétrique) générale de la loi de probabilité de cette population générale. Cette possibilité ne doit pas être négligée dans la mesure où la connaissance de la forme générale de la loi de probabilité permet de mieux choisir la méthode d'estimation statistique des paramètres de cette population ainsi que la méthode de traitement statistique du fichier. On sait que l'analyste s'assure par des *tests d'ajustement* (cf. § 10.3 et 11.1) que les hypothèses qu'il a avancées sur la forme générale de la loi de probabilité des observations analysées ne contredisent pas la nature et la spécificité des données initiales.

Etape 5 : composition d'un plan détaillé de l'analyse numérique du fichier. Cette étape commence par la composition d'un dossier statistique sur les données recueillies et sur les résultats de l'analyse préliminaire. On définit les principaux groupes à analyser. On procède à une mise à jour du dictionnaire des notions. On décrit soigneusement l'organigramme de l'analyse en indiquant les méthodes sollicitées. *On formule le critère d'optimisation* qui présidera au choix de la méthode (ou de la famille de méthodes) de traitement statistique du fichier (cf. § 1.2).

Etape 6 : réalisation numérique de la principale partie du traitement statistique du fichier. Le principal souci de l'analyste à cette étape est la conduite efficace des calculs moyennant une formulation adéquate du problème de traitement et de description des données initiales dans le langage d'entrée du paquet. Il faut tenir compte de la dimension du problème, de la complexité de l'algorithme, des performances de l'ordinateur (longueur des mots, rapidité, capacité de la mémoire, organisation du fichier, etc.) et enfin des particulari-

tés des données (degré de conditionnement des matrices utilisées pour la réalisation des procédures linéaires, fiabilité des estimations *a priori* des paramètres, etc.).

Etape 7 : bilan de l'analyse. Cette étape débute par la rédaction d'un compte rendu statistique formel sur l'analyse effectuée. L'interprétation des résultats des procédures statistiques (estimation des paramètres, tests entre hypothèses, applications dans un espace de dimension moindre, classification, etc.) tient compte aussi bien de la place de ces procédures dans l'organigramme que du rapport des tailles des échantillons utilisés, de la dimension de l'espace des observations, du nombre et des valeurs des paramètres. L'aspect théorique de ces questions a été peu exploré malgré leur extrême actualité. Font exception les travaux [42], [99], [100]. Dans les cas où la théorie est incapable d'interpréter les résultats des calculs, on peut appeler à la rescousse la simulation statistique (cf. §§ 3.3 et 6.3).

Ensuite les résultats de l'analyse et les principales conclusions sont présentés dans les termes de la discipline utilisatrice. Si l'analyse a été conduite avec des méthodes et des modèles de statistique mathématique, les conclusions sont formulées en termes d'estimations des paramètres inconnus du système analysé ou sous forme de réponse à la question de la validité de l'hypothèse statistique vérifiée et sont accompagnées des estimations du degré de leur validité. En revanche si l'étude est réalisée par les méthodes de l'analyse des données (c'est-à-dire dans le cadre de la deuxième approche), les conclusions ne peuvent être interprétées probabilistiquement.

En conclusion on vérifie dans quelle mesure les objectifs définis dans l'étape 1 ont été atteints, et si certains ne l'ont pas été, on explique pour quelle raison. Le travail s'achève par la formulation en termes adéquats des nouveaux problèmes issus de l'analyse effectuée.

Dans certains traités de théorie générale de statistique (cf. par exemple [131]) les étapes 5, 6 et 7 sont regroupées en une seule intitulée « Traitement et analyse ».

Concluons la description du schéma logique général de l'analyse des données initiales en signalant que les méthodes de simulation statistique et les méthodes de traitement statistique primaire sont transcendantes dans la réalisation des importantes étapes 1, 4 et 7 et en cas de nécessité peuvent être appliquées aux étapes 3, 5 et 6.

1.2. Formulation d'optimisation des principaux problèmes de statistique appliquée et stabilité des inférences statistiques

1.2.1. Lien entre la formulation d'optimisation des principaux problèmes de statistique appliquée et la stabilité des inférences statistiques. Nous avons signalé plus haut que les principaux problèmes de traitement statistique du fichier pouvaient être énoncés *sous*

la forme d'un problème d'optimisation général (avec un critère d'optimisation de la qualité de la méthode choisi en conséquence) de telle sorte que et les méthodes de mathématique statistique et les méthodes logico-algébriques soient élaborées comme solutions de ce problème. Montrons que si le critère d'optimisation de la qualité de la méthode et la classe de solutions admissibles sont dûment choisis, on peut dans le cadre d'une même approche obtenir des méthodes et des modèles éprouvés dans les problèmes

d'étude statistique des dépendances;

de classification des objets ou des variables;

de représentation condensée des données.

Par ailleurs, l'introduction de critères de qualité d'une méthode (n° 1.1.2) permet de mettre en œuvre une technique utile de recherche de méthodes de traitement statistique conduisant à des inférences *stables* par rapport à une variation des conditions initiales (par rapport à la nature et la précision de l'enregistrement des données traitées). Il faut, en particulier, résoudre *plusieurs fois* un problème d'optimisation afin de déterminer la meilleure méthode de traitement statistique des données pour divers critères de qualité, par exemple, pour des critères formant toute une famille dépendant d'un ou de plusieurs paramètres. Ceci nous conduit en définitive à un ensemble d'inférences statistiques : à chaque critère correspond sa meilleure méthode, à chaque meilleure méthode, son inférence statistique. Dans cet ensemble d'inférences statistiques il faut donc en choisir une ou plusieurs qui varient peu quand on passe d'un critère à un autre dans un domaine de variation assez vaste [8], [11], [137].

L'adéquation et l'actualité d'une telle procédure de traitement statistique des données sont dues au fait que, en général, ni l'information initiale sur la nature probabiliste des données à traiter ni la connaissance du mécanisme « physique » de l'événement étudié ne nous fournissent des motifs suffisants susceptibles de justifier le choix de *tel modèle* et, partant, *de tel critère de qualité de la méthode.* Or cela signifie qu'il faut tenir en réserve *toute une classe* de modèles admissibles (de critères). Ceci explique pourquoi les inférences statistiques fondées sur le principe du maximum de vraisemblance pèchent souvent par la stabilité de leurs « bonnes » propriétés (la réalisation de ce principe implique que soit postulé *a priori* un type de loi de probabilité des données traitées).

De façon plus concrète, le § 10.4 développe une forme d'établissement d'inférences statistiques stables pour un problème d'estimation.

1.2.2. Etude statistique des dépendances des variables. L'étude du caractère et de la structure des dépendances des variables caractérisant l'état ou le comportement des objets (processus) sondés est le principal objectif de l'analyse statistique multidimensionnelle. Rien

d'étonnant donc que le problème qui sert d'intitulé à ce numéro prévale incontestablement tant du point de vue de son actualité que du point de vue de la richesse et du degré d'élaboration de l'appareil mathématique correspondant. Font partie de ce dernier les méthodes d'analyse de régression, d'analyse des corrélations, des variances et covariances, de planification extrémale des expériences de régression, les méthodes d'analyse des séries temporelles, certains méthodes et modèles de dépendances d'un type spécial (par exemple, markovien).

Dans la plupart des cas, le schéma général d'étude de ce problème est le suivant. Le vecteur X des variables enregistrées est décomposé en deux sous-vecteurs dont l'un, par exemple,

$$X^{(1)} = \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{pmatrix},$$

est interprété comme le vecteur des caractéristiques *des conditions de fonctionnement* (ou *d'état*) du système étudié (en principe, tout ou partie de ces caractéristiques font l'objet d'une régulation ou d'une commande partielle) et le second,

$$X^{(2)} = \begin{pmatrix} x^{(m+1)} \\ \vdots \\ x^{(p)} \end{pmatrix},$$

comme un vecteur des *variables résultantes* caractérisant le *comportement* ou l'*efficacité du fonctionnement* (la *qualité*) du système analysé*). Le problème consiste à expliquer de façon suggestive le comportement des variables résultantes $X^{(2)}$ par une variation des variables d'entrée $X^{(1)}$, c'est-à-dire à définir une fonction vectorielle

$$\mathbf{f}(X^{(1)}) = \begin{pmatrix} f_1(X^{(1)}) \\ \vdots \\ f_{p-m}(X^{(1)}) \end{pmatrix}$$

dans la classe des solutions admissibles **F**, qui donne la meilleure approximation, dans un certain sens, du comportement du vecteur $X^{(2)}$ sur l'ensemble des points-observations $\{X_i\}_{i=1,\ldots,n}$. Pour formuler mathématiquement ce problème, introduisons les quantités $\delta_{ki} = \delta(f_k(X_i^{(1)}))$ qui caractérisent les erreurs affectant la description de la variable résultante $x^{(m+k)}$ par la fonction $f_k(X^{(1)})$ au

*) Les variables $X^{(1)}$ sont souvent appelées variables *d'entrée* ou *explicatives*, ou encore prédicteurs (en économétrie la plupart d'entre elles sont des *variables exogènes*). Les variables résultantes $X^{(2)}$ s'appellent variables *dépendantes* ou *prédictantes*, ou *réponses* (en économétrie la plupart d'entre elles forment les *variables endogènes*).

point X_i, puis la fonctionnelle

$$\Delta (\mathbf{f}) = \Delta (\{\delta_{ki}\}, \quad k = 1, \ldots, p - m, \quad i = 1, \ldots, n) \tag{1.6}$$

comme une mesure de l'adéquation du modèle

$$X^{(2)} \approx \mathbf{f} (X^{(1)}).$$

L'étude statistique des dépendances des variables $X^{(2)}$ par rapport aux variables $X^{(1)}$ revient donc à déterminer une fonction vectorielle $\hat{\mathbf{f}} (X^{(1)})$ qui soit solution du problème d'extremum

$$\Delta (\hat{\mathbf{f}} (X^{(1)})) = \operatorname*{extr}_{\mathbf{f} \in \mathbf{F}} \Delta (\mathbf{f} (X^{(1)})). \tag{1.7}$$

La forme concrète des erreurs δ_{ki}, de la fonctionnelle d'adéquation Δ (f) et de la classe de fonctions admissibles **F** dépend de la nature des données initiales analysées et de certaines informations initiales (sous réserve qu'elles existent) sur la nature et la structure des dépendances cherchées. Si **F** est une famille $\{\mathbf{f} (X^{(1)}; \Theta)\}$ de fonctions dépendant d'un paramètre, le problème (1.7) se réduit à la détermination (estimation statistique) des valeurs des paramètres $\hat{\Theta}$ réalisant l'extremum (1.6) et les méthodes correspondantes sont dites méthodes *paramétriques*.

Dans [10] on trouvera les variantes de ce schéma général qui sont les plus couramment utilisées en théorie et en pratique. Ces variantes étudient presque toutes des modèles additifs d'approximation-régression de la forme

$$X^{(2)} = \mathbf{f} (X^{(1)}) + \varepsilon, \tag{1.8}$$

dans lesquels le vecteur colonne à $(p - m)$ dimensions des restes ε décrit soit (à titre de composante aléatoire résiduelle) l'influence de l'ensemble des facteurs aléatoires négligés sur $X^{(2)}$, soit (à titre d'erreur d'approximation) le degré d'approximation des variables $X^{(2)}$ par des fonctions de classe **F**, soit (chose plus fréquente en réalité) l'un et l'autre à la fois.

Il est évident que la variante paramétrique du modèle (1.8) peut être mise sous la forme

$$X^{(2)} = \mathbf{f} (X^{(1)}; \Theta) + \varepsilon. \tag{1.8'}$$

Selon le caractère des conditions supplémentaires imposées à la nature des restes ε et à la classe de fonctions **F**, on débouche sur telle ou telle forme concrète des erreurs δ_{ki} et de la fonctionnelle Δ, ce qui définit le type des modèles d'approximation-régression et la méthode d'estimation des paramètres inconnus du modèle.

La méthode de description formalisée du problème d'étude statistique des dépendances quoique assez générale n'a pas l'ambition

de couvrir l'ensemble de toutes les positions des problèmes et modèles se rapportant à ce sujet.

1.2.3. Classification des objets ou des variables. S'agissant de la classification d'un ensemble *d'objets*, on conviendra que chacun d'eux est défini par une colonne de la matrice (1.4) ou bien que la structure géométrique de leurs distances (liens) est donnée par la matrice (1.4'). Dans la classification *des variables*, l'information initiale admet la même interprétation sauf que chaque variable est donnée par une ligne de la matrice (1.4). Dans la suite, si le contraire n'est pas spécialement mentionné, nous ne ferons pas de distinction entre les objets et les variables en raison de l'identité de la position des problèmes et de la principale méthode d'étude.

La position générale (non rigoureuse) du problème de classification des objets consiste à subdiviser la population des objets $\mathbf{O} = \{O_i\}_{i=1, \ldots, n}$ représentée sous la forme (1.4) ou (1.4') en un petit nombre de groupes ou classes homogènes dans un certain sens. Pour formaliser ce problème, il est commode d'assimiler les objets analysés aux points de l'espace factoriel : si les données initiales sont de la forme (1.4), ces points sont la représentation géométrique d'observations multidimensionnelles $X_1, X_2, \ldots, X_n$ dans un espace $\mathbf{X}$ à p dimensions ramené aux axes de coordonnées $0x^{(1)}, 0x^{(2)}, \ldots, 0x^{(p)}$; si les données initiales sont de la forme (1.4'), l'analyste ne connaît pas les coordonnées de ces points, mais la structure de leurs distances. Il est naturel de postuler que la proximité géométrique de deux ou plusieurs points de cet espace est synonyme de proximité des états « physiques » des objets correspondants, de leur homogénéité. La classification revient alors à diviser la population des points-observations en un petit nombre — connu ou pas — de clusters (agglomérats, taxons, formes) situés à une certaine distance l'un de l'autre (au sens de la métrique définie dans l'espace $\mathbf{X}$ correspondant) mais qui ne se partitionnent pas eux-mêmes en classes aussi éloignées *).

Il est évident que l'algorithme de classification S doit obéir à certaines règles, règles qui peuvent être formulées dans un cas assez général sous forme d'un critère ou d'une fonctionnelle de qualité de la classification $Q(S)$. La forme de cette fonctionnelle, la position du problème de classification et la définition de la classe $\mathbf{A}$ des règles admissibles de classification dépendent du caractère de l'informa-

*) La structure géométrique de la population des objets peut être telle que cette classification ne passe pas (par exemple, si les points sont uniformément distribués dans un domaine de l'espace $\mathbf{X}$). L'application formelle de ces méthodes de classification à de telles populations conduit dans ces cas à envisager une ordinaire partition du domaine étudié de l'espace $\mathbf{X}$ en un certain nombre de sous-domaines de groupement (analogues multidimensionnels des intervalles de groupement obtenus lors du traitement statistique primaire des observations en dimension un).

tion initiale sur les classes cherchées et de la présence (ou de l'absence) d'une information préliminaire empirique (dite échantillon d'apprentissage) sur ces classes.

Ainsi, dans le cas général le problème de classification d'une population O d'objets représentée sous la forme (1.4) ou (1.4') peut être formulé comme le problème de la détermination d'une partition (règle de classification) S^* de la population O en classes disjointes $S_1^*, S_2^*, \ldots, S_k^*$; $\bigcup_{j=1}^{k} S_j^* = \mathrm{O}$, $S_i \cap S_j = \varnothing$ pour $i \neq j$, telle que la fonctionnelle de qualité $Q(S)$ soit minimale, c'est-à-dire

$$Q(S^*) = \min_{S \in \mathrm{A}} Q(S),$$

ou maximale, c'est-à-dire

$$Q(S^*) = \max_{S \in \mathrm{A}} Q(S).$$

Le nombre de classes k peut être inconnu ou donné.

Dans [10] est exhibée la forme concrète des fonctionnelles $Q(S)$ et des classes A des règles de résolution admissibles conduisant à des schémas connus tels que l'*analyse discriminante*, la *décomposition des mélanges des lois de répartition*, l'*analyse des clusters*, etc.

1.2.4. Abaissement de la dimension de l'espace factoriel et sélection des variables les plus informatives. Le passage d'un nombre élevé p de variables initiales caractérisant l'état (le comportement, l'efficacité du fonctionnement) du système analysé à un nombre p' plus petit de variables plus informatives (qui soit sont sélectionnées selon un certain critère parmi les variables initiales, soit en sont des fonctions) est dicté au moins par trois raisons essentielles: *primo*, par le *recouvrement de l'information* fournie par des variables fortement corrélées; *secundo*, par la *non-informativité des variables variant peu* au passage d'un objet à un autre (faible variabilité des variables); *tertio*, par l'*éventualité d'une agrégation*, c'est-à-dire d'une sommation simple ou pondérée, de certaines variables. Formellement, le passage (avec une perte minimale en information) à une nouvelle collection de variables $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ peut être décrit de la manière suivante. Soit $Z = Z(X)$ une fonction vectorielle p'-dimensionnelle des variables initiales $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ $(p' < p)$ et soit $I_{p'}(Z(X))$ la mesure d'informativité du système p'-dimensionnel de variables $Z(X) = (z^{(1)}(X), z^{(2)}(X), \ldots, z^{(p')}(X))$, mesure qui est donnée *a priori*. Le choix de la fonctionnelle $I_{p'}(Z)$ est tributaire de la spécificité du problème réel résolu et repose sur l'un des deux critères suivants: un critère d'*autoinformativité* visant à conserver dans le fichier $\{X_i\}_{i=1,\ldots,n}$ le maximum d'information fournie par les variables initiales et un critère d'*informativité exogène* visant à «presser» de $\{X_i\}_{i=1,\ldots,n}$ le maximum d'information fournie par d'autres variables (exogènes).

Le problème consiste à définir une collection de variables $\tilde{Z}$ appartenant à une classe F de transformations admissibles des variables initiales $x^{(1)}, \ldots, x^{(p)}$ telle que

$$I_{p'}(\tilde{Z}(X)) = \max_{Z \in F} I_{p'}(Z(X)).$$

Les diverses variantes de concrétisation de cette position (qui dictent le choix de la mesure d'informativité $I_{p'}(Z)$ et de la classe F des transformations admissibles) conduisent à des méthodes concrètes d'abaissement de la dimension : *méthode des composantes principales*, *analyse factorielle*, *groupement extrémal des paramètres*, etc. (cf. [10]).

Conclusions

1. La *statistique appliquée* est une discipline scientifique qui élabore et systématise les notions, méthodes et modèles mathématiques destinés à l'organisation de la collecte, de l'enregistrement standard, de la systématisation et du traitement (y compris avec un ordinateur) des données statistiques en vue d'une représentation commode, d'une interprétation et de l'établissement de conclusions scientifiques et pratiques.

2. Les problèmes de l'élaboration des méthodes de définition d'un système de variables initiales caractérisant l'objet ou le processus étudié relèvent des *disciplines utilisatrices* (économie, médecine, etc.).

3. La *statistique mathématique* qui conçoit et baille une partie de l'appareil mathématique utilisé par la statistique appliquée est complètement déchargée de fonctions propres à cette dernière telles que

l'ajustement et la mise au point des instruments mathématiques en fonction des traits spécifiques du problème réel étudié ;

l'élaboration de techniques logico-algébriques de traitement statistique des données, c'est-à-dire de techniques ne conjecturant pas la nature probabiliste des données traitées ;

la réduction des diverses formes d'information à une forme standard de représentation des données initiales et leur préparation au traitement ;

l'organisation d'un traitement automatique sur ordinateur, la mise en place d'un logiciel.

4. Il est commode de diviser la procédure d'analyse statistique des données en quatre *étapes essentielles* :

étape 1 : analyse initiale (préalable) du système étudié ;

étape 2 : établissement d'un plan de collecte de l'information initiale ;

étape 3 : collecte des données initiales, préparation et introduction de ces données dans l'ordinateur ;

étape 4 : traitement statistique préliminaire des données ;

étape 5 : sélection des principaux méthodes et algorithmes de traitement statistique des données, établissement d'un plan circonstancié d'analyse numérique des données ;

étape 6 : réalisation du plan d'analyse numérique des données initiales (exploitation directe de l'ordinateur) ;

étape 7 : bilan de l'analyse.

Les étapes ont été énumérées dans l'ordre chronologique, mais, en cas de nécessité, elles peuvent être réalisées par une procédure séquentielle : les résultats de la réalisation d'étapes en aval peuvent nécessiter le retour à des étapes en amont avec une prise en considération de la nouvelle information.

5. Chaque problème fondamental de statistique appliquée peut être énoncé *sous forme d'un problème d'optimisation général* pour un certain critère d'optimisation (une fonctionnelle) de la qualité de la méthode. Cette présentation en termes d'optimisation permet notamment :

de définir le cadre dans lequel il faut chercher les méthodes probabilistes et logico-algébriques de traitement statistique des données ;

de décrire une méthode constructive d'établissement d'inférences statistiques stables.

6. Le principal objectif de l'analyse statistique multidimensionnelle est l'étude des dépendances entre les variables pour résoudre les problèmes suivants : reconstitution de la valeur d'une variable résultante à partir des valeurs des variables « illustratives » ; la classification des observations multidimensionnelles ; l'abaissement de la dimension de l'espace factoriel étudié.

CHAPITRE 2

MÉTHODE DE RAISONNEMENT PROBABILISTE EN STATISTIQUE APPLIQUÉE

2.1. Théorie des probabilités et conditions de son application

2.1.1. Ensemble statistique et « jeu du hasard ». Au chapitre 1 nous avons signalé l'existence de deux techniques de traitement statistique des données: une technique mathématico-statistique (probabiliste) et une technique logico-algébrique. L'analyste n'est *contraint* de recourir à la deuxième approche que lorsque les conditions de collecte (d'enregistrement) des données initiales ne s'inscrivent pas dans l'*ensemble statistique*, c'est-à-dire dans les cas où il n'existe aucune possibilité pratique ou à la rigueur virtuelle de reproduire fidèlement les principales conditions dans lesquelles se sont déroulées les mesures des données analysées.

Dans les conditions d'ensemble statistique l'analyste a la possibilité d'appliquer les méthodes mathématico-statistiques classiques de traitement des données pour justifier le meilleur choix des méthodes de traitement, de représentation finale et d'interprétation des données analysées lorsqu'il utilise les diverses informations initiales sur leur nature *aléatoire* (*stochastique*). Ceci étant, il est entendu que même la fidélité — que nous avons postulée — de la reproduction des principales conditions de l'expérience ou de l'observation *ne nous met pas à l'abri* dans la plupart des situations réelles (compte tenu de leur complexité, de leur multiplicité et de l'étude incomplète des facteurs qui en sont la cause) *d'une dispersion aléatoire* des résultats de l'observation. Ainsi, une machine-transfert même réglée de façon idéale n'est pas en mesure de produire des pièces absolument identiques entre elles (et au modèle). On est confronté aux mêmes conditions d'ensemble statistique et de dispersion aléatoire des variables étudiées quand on analyse le nombre de pièces défectueuses dans un lot d'une certaine taille prélevé dans une production en série en régime stationnaire ou bien le revenu moyen par personne dans des foyers pris au sein d'une population homogène (sous un angle social, géographique, économique, etc.).

Mais partout où le hasard semble se jouer à la surface, il est toujours sous l'empire de lois internes cachées. Or, ce sont justement la *théorie des probabilités* et la *statistique mathématique* qui offrent à l'analyste et des modèles mathématiques différents de telles lois internes cachées et les outils indispensables à l'analyse théorique et empirique de leurs propriétés et corrélations. Les parties II et III sont consacrées à la description du strict minimum de leurs notions fondamentales.

2.1.2. Théorie des probabilités et conditions d'ensemble statistique. « La théorie des probabilités est une science qui permet à partir des probabilités d'événements aléatoires de déterminer les probabilités d'autres événements aléatoires liés d'une certaine manière aux premiers... On peut dire aussi que la théorie des probabilités est une science mathématique définissant (en termes de modèles mathématiques ! — *S.A.*) les lois régissant l'interaction d'un grand nombre de facteurs aléatoires » *).

On a déjà signalé que la méthode probabiliste de raisonnement (et l'appareil mathématique correspondant) se prêtait le mieux à l'étude des problèmes soumis à l'action d'un *système de conditions stationnaire* (c'est-à-dire ne dépendant pas du temps) comprenant un grand nombre de facteurs incontrôlables qui ne permettent pas de dire avec certitude si l'événement qui nous intéresse aura lieu ou non. Ceci étant, on admet qu'on a en principe la possibilité (au moins virtuelle) de répéter plusieurs fois l'expérience ou l'observation effectuée dans le cadre des mêmes conditions. Ce sont ces conditions qu'on qualifie ordinairement de *conditions d'action d'un ensemble statistique* ou encore de conditions de respect de l'homogénéité statistique de la population analysée.

Les exemples les plus simples et les plus suggestifs de situations réelles vérifiant la condition de stabilité statistique (ou s'inscrivant dans le cadre d'un ensemble statistique) nous sont fournis par les jeux de hasard **). En effet, quand on jette une pièce, un dé ou quand on tire une carte d'un paquet, et qu'on s'intéresse à la probabilité d'apparition de « face », du « six » ou de la « dame de pique », on a toutes les raisons de supposer que :

a) l'on peut répéter plusieurs fois la même opération dans les mêmes conditions ;

b) la présence d'un grand nombre de facteurs aléatoires caractérisant les conditions de réalisation de chaque opération ne permet pas d'affirmer que l'événement nous intéressant aura bien lieu ou non ;

c) plus le nombre d'expériences sera élevé et plus les fréquences relatives d'apparition des événements étudiés seront proches de certaines valeurs constantes appelées *probabilités* de ces événements, plus exactement, la fréquence relative d'apparition de « face » sera proche de $^1/_2$, celle du « six », proche de $^1/_6$ et celle de la « dame de pique », proche de $^1/_{52}$ (pour un jeu de 52 cartes).

*) *Encyclopédie mathématique*. M., Sovietskaïa entsiklopedia, 1976, t. I, p. 655-656.

**) Historiquement les jeux de hasard sont à l'origine de l'émergence et de l'intense développement de la théorie des probabilités. Les premiers résultats intéressants sont l'œuvre de L. Pacioli (« Summa de arithmetica, geometria, proportioni e proportionalita », 1494, Venezia), de G. Cardano (« Opera Omnia », 1563, tome 1) et de N. Tartaglia (« General Trattato di numeri e misure », Venezia, 1560). Cf. [6].

Il est évident que les conditions d'ensemble statistique impliquent dans le cas des trois expériences précédentes que la pièce de monnaie soit symétrique (ou que la pièce de rechange soit absolument la même), que le dé ne soit pas pipé et que la carte tirée soit remise dans le paquet qui sera soigneusement battu avant le tirage suivant.

2.1.3. Classification des situations réelles par les conditions d'ensemble statistique. Dans les plus importantes sphères de l'activité humaine: économie, processus sociaux, technique, industrie, médecine, sciences, etc., il faut dans chaque cas concret s'assurer que les conditions d'ensemble statistique sont bien réunies.

La réalisation des propriétés a), b) et c) de stabilité statistique peut servir de critère à la classification des problèmes posés dans les divers domaines des connaissances humaines en trois catégories.

La *première catégorie* (ou *catégorie de haute performance des méthodes probabilistes et statistiques*) couvre les situations dans lesquelles les propriétés a), b) et c) de stabilité statistique sont indiscutablement réalisées ou bien sont si peu violées que cela n'affecte pratiquement pas les conclusions statistiques acquises à l'aide des modèles probabilistes. Outre la rubrique déjà signalée, cette catégorie englobe encore divers domaines d'*économie* et de *sociologie* et notamment les problèmes liés à l'étude du comportement d'un objet (un individu, une famille ou une autre unité socio-économique ou industrielle) en tant que représentant d'une grande population homogène composée d'objets identiques (cf. par exemple [3], [9], [137]). Un domaine de prédilection des méthodes probabilisto-statistiques est la *démographie* (cf. [32], [126]). Les notions probabilistes sont fondamentales dans les sciences de l'ingénieur telles que la *théorie de la fiabilité des systèmes* composés d'un grand nombre d'éléments (cf. [60]) et la *théorie du contrôle aléatoire de la qualité* d'une production (cf. [20], [60]). En *médecine*, les techniques probabilisto-statistiques ont permis d'introduire la notion des facteurs de risque de l'évolution des principales maladies chroniques et de quantifier leur influence contribuant ainsi à une plus grande individualisation, donc à une grande efficacité des mesures prophylactiques et des traitements (cf. [103], [135]). Les résultats d'études probabilisto-statistiques spéciales montrent que la probabilité de vivre jusqu'à un certain âge n'est pas affectée par d'importantes variations (en fonction des conditions de vie). Ces résultats ont servi de support à l'établissement de *tables de survie* (cf. [35]) qui, dans une certaine mesure et dans un certain sens (plus exactement du point de vue statistique et non individuel bien sûr !), bousculent la vieille maxime « personne ne connaît l'heure de sa mort ». Le mode de raisonnement probabilisto-statistique joue un rôle éminent dans les recherches en physique contemporaine (notamment, en *physique statistique*, cf. [74]) et en mécanique classique (en *théorie statistique des gaz*).

Signalons un trait général important propre à l'écrasante majorité des problèmes des activités humaines susmentionnées dans lesquelles les méthodes probabilisto-statistiques font largement recette. Il s'agit de la *très grande dimension* de l'information traitée, c'est-à-dire de situations dans lesquelles l'état ou le comportement de chacun de ces objets est décrit à tout instant *par tout un ensemble* de variables. Ces variables peuvent être aussi bien quantitatives (revenu moyen de chaque membre d'une famille, nombre d'une famille, volume de la production brute d'une entreprise, etc.) qu'ordinales (classification d'un spécialiste, caractéristique comparative des conditions de logement) ou nominales (profession, nationalité, sexe, causes de migration, etc.). Toutes ces variables sont liées entre elles de façon complexe.

Ces situations donnent lieu à des schémas multidimensionnels dont l'étude implique des méthodes d'*analyse statistique multidimensionnelle.*

La *deuxième catégorie* (ou catégorie des *applications probabilisto-statistiques admissibles*) englobe les situations caractérisées par d'importantes violations de l'*invariance des conditions de l'expérience* (deuxième moitié de la condition a)) et par les écarts par rapport à la condition c) qui en résultent. Une forme caractéristique des écarts par rapport aux conditions d'ensemble statistique est la réunion en une seule série d'observations (à traiter) de tranches de données initiales recueillies dans des conditions différentes (à des dates ou dans des populations différentes). Dans cette catégorie d'applications on peut aussi rapporter une classe de problèmes liés à l'analyse de *courtes séries temporelles* enregistrées dans des conditions excluant pratiquement la possibilité de mesurer simultanément plusieurs réalisations empiriques de la série temporelle étudiée sur un même intervalle de temps. L'usage des méthodes probabilisto-statistiques est possible dans ce cas, mais les résultats et conclusions sont sujets à caution (par exemple, il ne faut trop se fier aux caractéristiques numériques du degré de vraisemblance de ces conclusions, c'est-à-dire au *niveau de confiance*, au *seuil de signification* d'un test, etc.), aussi faut-il dans la mesure du possible les assister d'autres méthodes d'analyse scientifique.

La *troisième catégorie* enfin (ou catégorie des *applications probabilisto-statistiques inadmissibles*) comprend les situations caractérisées soit par le rejet de l'idée maîtresse de l'ensemble statistique, savoir la *multitude des unités* de la population étudiée (on entend par là l'impossibilité formulée dans la condition a) de répéter une même expérience dans les mêmes conditions), soit par le *déterminisme* total de l'événement étudié, c'est-à-dire l'absence de l'effet « perturbateur » des facteurs aléatoires (violation de la condition b)). Dans de telles circonstances l'analyste doit se servir de méthodes d'analyse des données (cf. [43]) et doit tirer un trait sur la possibilité d'in-

terpréter probabilistiquement les données traitées et les conclusions acquises par le traitement de ces données.

Il n'existe pas de méthodes mathématiques rigoureuses susceptibles de nous dire si nous sommes ou non dans des conditions d'ensemble statistique: tout modèle probabiliste de même que tout modèle mathématique n'est en général qu'une approximation de la réalité analysée. On ne peut envisager que les situations qui s'inscrivent manifestement dans un ensemble statistique (lancement d'une pièce de monnaie, d'un dé; contrôle des articles d'une production en série fonctionnant en régime stationnaire, etc.), celles qui s'y inscrivent approximativement, avec des réserves, et enfin celles qui sont complètement étrangères. Signalons que de nombreux points restent à éclaircir même dans la troisième catégorie. Ainsi, par exemple, du point de vue de l'ensemble statistique, les événements du type « en l'an 2000 une guerre éclatera entre les pays A et B » ne font pas partie de la deuxième catégorie à cause de la violation de la condition a)! Mais il existe une conception des *probabilités subjectives* (cf. par exemple [90]) qui permet d'envisager de tels événements. Pour cela il suffit de solliciter l'avis d'experts et au lieu de répéter plusieurs fois une expérience dans les mêmes conditions on peut se contenter de l'analyse de la situation par des experts. Il est évident que l'*expert est alors interprété comme un instrument de mesure* fournissant des résultats entachés d'une erreur aléatoire. La précision de cet « instrument de mesure », c'est-à-dire la précision de la « prédiction » par l'expert de l'événement analysé, dépend aussi bien de l'influence objective de facteurs aléatoires « perturbateurs » (c'est-à-dire de l'échéance de l'événement, de la complexité de la situation, etc.) que de la documentation, de la clairvoyance et de bien d'autres qualités de l'expert. Sans nous mêler du débat qui oppose les partisans de la probabilité subjective à ceux de la probabilité classique, disons seulement qu'en la matière le seul juge objectif est la pratique. Rappelons à ce propos la pensée développée par F. Engels dans l'« Anti-Düring »: « ... La mathématique, pourtant si austère, a commis le péché originel; elle a mangé le fruit de l'arbre de la science, qui lui a ouvert la carrière des progrès les plus gigantesques, mais aussi celle des erreurs. L'état virginal où tout ce qui était mathématique avait une valeur absolue et était démontré d'une manière irréfragable fut perdu pour jamais; alors s'ouvrit le règne des controverses, et nous en sommes arrivés au point que la plupart des gens différentient et intègrent sans comprendre ce qu'ils font, par un pur acte de foi, parce que cela a toujours réussi jusque-là » *).

En récapitulant le débat sur l'essence, la destination et les limites d'application de la théorie des probabilités, nous pouvons un siè-

*) F. Engels. *M.E. Düring bouleverse la science* (Anti-Düring). Traduit par Bracke (A.M. Desrousseaux). Tome 1, Philosophie, Paris, 1946, page 125.

cle et demi plus tard après le célèbre savant français Laplace répéter avec plus d'assurance et de conviction qu'il est remarquable que la science qui a commencé par l'étude des jeux se soit hissée aux plus importants objets de la connaissance humaine *).

2.2. « Rapports » entre la théorie des probabilités et la statistique mathématique

2.2.1. Mode statistique de prise de décision. Supposons que le lecteur observe deux personnes jouant au dé d'après les règles suivantes. Le joueur A reçoit un rouble de la part du joueur B si en quatre coups il sort au moins une fois le six (appelons cette issue « six ») et lui verse un rouble dans le cas contraire (appelons cette issue « non-six »). Après cent coups le lecteur doit remplacer l'un des deux joueurs avec la possibilité de miser soit sur l'issue « six », soit sur l'issue « non-six ». Comment doit-il faire correctement son choix ?

Le *mode statistique* de résolution de ce problème est dicté par le bon sens et consiste en ce qui suit. Après avoir suivi les cent coups des deux joueurs et calculé les fréquences relatives de leurs gains, le lecteur devrait en toute logique miser sur l'issue qui est apparue le plus souvent. Par exemple, le joueur B a gagné 52 parties sur 100, c'est-à-dire que le six n'est pas apparu une seule fois en 52 parties de quatre coups (réciproquement, il est apparu au moins une fois dans les 48 autres parties). Donc, si le lecteur applique le mode statistique de raisonnement, il juge qu'il aura intérêt à miser sur l'issue « non-six », c'est-à-dire sur l'issue dont la fréquence relative d'apparition ($\hat{p}$) est égale à 0,52 ($<$0,50).

2.2.2. Mode probabiliste de raisonnement. Ce mode est basé sur un *modèle mathématique* de l'événement étudié : si l'on admet que le dé est symétrique, donc que les chances d'apparition de chaque face sont égales (en d'autres termes, que la fréquence relative ou probabilité d'apparition du « un » est égale à celle du « deux », du « trois », . . ., du « six » et vaut $^1/_6$, on peut calculer la probabilité $p = P$ {« non-six »} de réalisation de l'événement « non-six », c'est-à-dire la probabilité de ne voir apparaître aucune fois le « six » en quatre coups successifs. Ce calcul est basé sur le modèle mathématique adopté. La probabilité de *non-apparition* du six *en un coup* de dé est égale à la somme des probabilités d'apparition du « un », du « deux », du « trois », du « quatre » et du « cinq », c'est-à-dire (en vertu de la définition de la probabilité d'un événement, cf. § 4.1) à $^5/_6$. On applique ensuite le théorème du produit des probabilités (cf. n° 4.1.3) qui dit que la probabilité d'événements indépendants est égale au produit des probabilités de chacun d'eux. Dans le cas

*) Œuvres complètes de Laplace. T. 7 : *Théorie analytique des probabilités*. Paris, Gauthier-Villars, 1886, p. CLII.

considéré, on s'intéresse à la réalisation de quatre événements indépendants consistant en la non-apparition du « six » et ayant la même probabilité de $^5/_6$. Donc

$$p = P\{\text{« non-six »}\} = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \frac{625}{1296} = 0{,}482.$$

On voit que la probabilité de l'événement « non-six » est inférieure à la demie, donc, celle de l'événement « six » est plus élevée : 1 — 0,482 = 0,518. Par conséquent, le lecteur qui aura choisi le mode probabiliste de raisonnement arrivera à une conclusion diamétralement opposée à celle du lecteur qui aura opté pour le mode statistique et il misera sur l'événement « six ».

2.2.3. Mode probabilisto-statistique (ou mathématico-statistique) de prise de décision. Ce mode est en quelque sorte une synthèse des deux précédents, puisqu'il utilise aussi bien les *données statistiques initiales* recueillies (sous forme des fréquences relatives d'apparition du « six » et du « non-six », qui, rappelons-le, sont respectivement égales à 0,48 et 0,52) par les observations que les *résultats fournis par le modèle probabiliste.* Mais le modèle adopté dans ce cas est plus souple, *il épouse la réalité en utilisant à cet effet l'information statistique accumulée.* En particulier, ce modèle ne postule plus la symétrie des dés, il admet que le centre de gravité soit déplacé. Le caractère de ce déplacement (si déplacement il y a) doit se manifester dans les données statistiques initiales recueillies. Mais le lecteur qui a choisi le mode probabilisto-statistique de raisonnement doit avoir présent à l'esprit que les valeurs des fréquences relatives des événements « six » et « non-six », déduites de ces données, ne sont que des *valeurs approchées* des vraies probabilités (des probabilités théoriques) de ces événements : en effet, quand on jette une pièce de monnaie idéalement équilibrée, disons 10 fois de suite, on peut très bien obtenir 7 piles ; donc, la fréquence relative de pile est égale à 0,7 ; mais cela ne signifie encore pas que les probabilités théoriques d'apparition de pile et de face sont respectivement égales à 0,7 et 0,3 : on sait que ces dernières sont égales à 0,5. De façon analogue, la probabilité de « non-six » observée dans les séries de cent coups (qui est égale à 0,52) peut être différente de la probabilité théorique, donc n'est pas une raison suffisante au choix de cette situation. Toute la question est de savoir de combien $\hat{p}_n$, probabilité observée (au terme de n épreuves), diffère de p probabilité théorique de l'événement qui nous intéresse, et comment l'écart $\hat{p}_n - p$ dépend du nombre n d'observations. (Intuitivement, il est clair que plus le nombre n d'observations est élevé et plus les fréquences relatives empiriques $\hat{p}_n$ sont crédibles, c'est-à-dire plus elles sont proches des probabilités théoriques p.) Dans notre cas, on obtient la ré-

ponse à cette question en raisonnant comme suit : a) on interprète la réalisation de tout nombre de parties comme une suite d'épreuves de Bernoulli, ce qui veut dire que le résultat de chaque partie est indépendant de celui des précédentes et que la probabilité inconnue p de l'événement « non-six » est la même pendant toutes les parties ; b) on utilise le fait que l'erreur $\Delta_n = \hat{p}_n - p$ suit approximativement la *loi normale* avec une moyenne nulle et une variance égale à $p(1-p)/n$ (cf. § 6.1).

Ces considérations permettent en particulier d'estimer la valeur absolue de l'erreur $|\Delta_n|$ commise en remplaçant la probabilité inconnue p de l'événement étudié (ici l'événement « non-six ») par la probabilité $\hat{p}_n$ calculée sur n épreuves (ici, $n = 100$ et $\hat{p}_{100} = 0{,}52$). Si l'on peut calculer la valeur absolue de l'erreur éventuelle Δ_n, il est naturel d'adopter la règle suivante de prise de décision : si la fréquence relative $\hat{p}_n$ de l'événement « non-six » est supérieure à 0,5 et continue de l'être après soustraction de l'erreur éventuelle $|\Delta_n|$, l'observateur a alors intérêt à miser sur « non-six » ; si $\hat{p}_n$ est inférieure à 0,5 et continue à l'être après addition de l'erreur éventuelle $|\Delta_n|$, il a intérêt à miser sur l'événement « six » ; dans les autres cas, il n'a aucune raison, sur le plan statistique, de préférer un événement à l'autre (c'est-à-dire qu'il doit ou bien poursuivre les observations, ou bien miser sur l'un ou l'autre des deux événements sachant pertinemment que son gain ou sa perte ne seront pas élevés).

Le calcul approché de la plus grande valeur de cette erreur d'après le théorème de Moivre-Laplace (cf. § 7.3) nous montre que, dans ce cas, on a presque sûrement (plus exactement avec une probabilité de 0,95) l'inégalité

$$|\hat{p}_n - p| \leqslant 2\sqrt{\frac{p(1-p)}{n}}. \tag{2.1}$$

En élevant (2.1) au carré et en résolvant l'inéquation obtenue par rapport au paramètre p, on trouve

$$\frac{\hat{p}_n + \frac{2}{n}}{1+\frac{4}{n}} - \frac{2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{1}{n^2}}}{1+\frac{4}{n}} \leqslant p \leqslant$$

$$\leqslant \frac{\hat{p}_n + \frac{2}{n}}{1+\frac{4}{n}} + \frac{2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{1}{n^2}}}{1+\frac{4}{n}}$$

ou, à $O(1/\sqrt{n})$ près,

$$\hat{p}_n - 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leqslant p \leqslant \hat{p}_n + 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}. \qquad (2.2)$$

Dans le cas considéré (pour $\hat{p}_n = 0{,}52$ et $n = 100$), on obtient

$$|\Delta_n| \approx 2\sqrt{\hat{p}_n(1-\hat{p}_n)}/\sqrt{n} = 2\sqrt{0.52\cdot(1-0{,}52)}/\sqrt{100} \approx 0{,}10.$$

Donc,

$$0{,}52 - 0{,}10 \leqslant p \leqslant 0{,}52 + 0{,}10.$$

Donc, les observations des issues des cent parties nous permettent seulement de conclure que la probabilité de l'événement « non-six » peut être un nombre quelconque de l'intervalle [0,42; 0,62], c'est-à-dire qu'elle peut être aussi bien inférieure à 0,5 (auquel cas il faut miser sur « six ») que supérieure à 0,5 (et il faut alors miser sur « non-six »).

En d'autres termes, le lecteur qui utilise le mode probabilisto-statistique de résolution du problème aboutit à une conclusion plus prudente: les cent parties qui lui ont servi de données statistiques initiales ne suffisent pas à définir avec certitude la plus probable des issues. D'où la solution: soit poursuivre son rôle de «spectateur» jusqu'à ce que le domaine de valeurs des probabilités p déduites de la double inégalité (2.2) ne soit entièrement situé à gauche ou à droite de 0,5, soit participer au jeu en le jugeant proche d'un jeu « inoffensif », c'est-à-dire un jeu dans lequel il pourra sauver sa mise dans une longue série de parties.

Cet exemple illustre le rôle et la vocation des méthodes probabilistes et de statistique mathématique ainsi que les liens qui les unissent. Si la *théorie des probabilités offre à l'analyste un ensemble de modèles mathématiques* destinés à décrire les lois régissant des événements ou le comportement de systèmes fonctionnant sous l'effet d'un grand nombre de facteurs aléatoires corrélés, la *statistique mathématique permet de sélectionner dans un ensemble de modèles probabilistes celui qui correspond le mieux, dans un certain sens, aux données statistiques dont dispose l'analyste.*

Conclusions

1. La *théorie des probabilités* est une discipline mathématique visant à élaborer (et à étudier les propriétés) de modèles mathématiques simulant les mécanismes de fonctionnement d'événements réels ou de systèmes dont les conditions de « vie » sont perturbées par un grand nombre de facteurs aléatoires (c'est-à-dire échappant à un contrôle strict).

2. La *statistique mathématique* est un système de notions, recettes et méthodes mathématiques utilisant des modèles probabilistes et destinées à la collecte, à la systématisation, à l'interprétation et au traitement des données statistiques pour en tirer des conclusions d'ordre scientifique et pratique. L'une des principales vocations des méthodes de mathématique statistique est le choix, dans un ensemble de modèles probabilistes admissibles, du modèle qui répond le mieux aux données statistiques caractérisant le comportement réel du système étudié.

3. Le *domaine d'application des méthodes probabilisto-statistiques* est défini à strictement parler par le respect (au moins approximativement) des conditions d'ensemble statistique, plus exactement : a) par la possibilité (virtuelle) de répéter les expériences ou les observations dans les mêmes conditions ; b) par la présence d'un grand nombre de facteurs aléatoires caractérisant les conditions de réalisation des expériences (des observations) et ne permettant pas de conclure à la réalisation ou non de l'événement qui nous intéresse.

4. *Il n'existe pas de méthodes mathématiques rigoureuses indiquant que l'on se trouve dans les conditions d'ensemble statistique* : tout modèle probabiliste de même que tout modèle mathématique n'est qu'une approximation de la réalité étudiée. Les situations réelles étudiées peuvent être, comme nous l'avons déjà signalé, classées conventionnellement dans trois catégories : 1) la catégorie de haute performance des méthodes probabilisto-statistiques ; 2) la catégorie des applications probabilisto-statistiques admissibles ; 3) la catégorie des applications probabilisto-statistiques inadmissibles. A noter que bien des situations classées dans la troisième catégorie peuvent être traitées aussi bien par les méthodes d'analyse des données que par des méthodes statistiques utilisant les probabilités subjectives.

CHAPITRE 3

MODÈLES MATHÉMATIQUES EN STATISTIQUE APPLIQUÉE

3.1. A quoi servent les modèles mathématiques ?

§3.1.1. Deux façons de voir la modélisation statistique. En étudiant au chapitre 2 les rapports entre la théorie des probabilités et la statistique mathématique, nous avons été amenés à conclure que si la théorie des probabilités livre à l'analyste un ensemble de *modèles mathématiques* imitant les mécanismes de fonctionnement de phénomènes réels *hypothétiques* ou de systèmes de nature stochastique, l'une des principales vocations de la statistique mathématique est le *choix fondé*, au sein d'un ensemble de modèles admissibles, du modèle qui répond le mieux (dans un certain sens) aux données statistiques initiales caractérisant le comportement réel du système analysé.

Donc, le meilleur traitement statistique des données initiales est un problème dont la résolution dépend en premier lieu de la *connaissance* de modèles convenables et de l'*habileté* de l'analyste à les « ajuster » à la réalité étudiée, et au besoin, à *construire un nouveau modèle* reflétant les traits spécifiques du problème étudié. Ce chapitre est consacré à l'exposition de certaines notions et généralités relatives à cette « connaissance » et cette « habileté ».

La construction et l'expérimentation d'un modèle, c'est-à-dire la description mathématique des relations entre les éléments réels du système analysé, impliquent généralement l'utilisation simultanée de deux types d'information : a) une *information a priori* sur la nature et le caractère des relations analysées ; b) des *données statistiques initiales* caractérisant le processus et le résultat du fonctionnement du système analysé. Ceci étant, on opte pour l'une des deux approches (plus exactement, pour la première ou pour leur combinaison). Si l'analyste dispose des deux types d'information, alors il se sert généralement d'une méthode de modélisation mathématique pour décrire les relations cherchées par des *équations analytiques* en s'appuyant sur l'information initiale (formalisée mathématiquement sous forme d'hypothèses initiales) caractérisant ces relations, puis par un traitement statistique de l'information b) estimer les valeurs numériques des paramètres figurant dans ces équations (étape d'ajustage du modèle). Si en revanche l'analyste ne dispose que d'une information initiale de type a) ou d'une information des deux types et qu'il veuille simuler le comportement du système analysé sous une

variation des paramètres figurant dans la description analytique du modèle ou de générer artificiellement (en s'appuyant sur les relations décrivant le modèle) des données statistiques de type b) pour les compléter, alors il doit en plus des éléments de modélisation mathématique mentionnée ci-dessus s'assurer le concours d'un ordinateur. Ce type de modélisation s'appelle *modélisation statistique* ou *de Monte Carlo.*

3.1.2. Notion de modèle mathématique. Le modèle mathématique est une abstraction du monde réel dans laquelle les relations entre les éléments réels, qui intéressent l'analyste, sont remplacées par des relations convenables entre des objets mathématiques que nous appellerons *relations de fonctionnement* du modèle. Les modèles mathématiques décrits par des variables aléatoires sont dits *probabilistes* ou *stochastiques.* Tout modèle est une représentation simplifiée de la réalité et tout l'art de la simulation est de savoir quoi, où, quand et comment simplifier.

L'exemple suivant permettra au lecteur de « prendre la mesure » de certains moments clés et de certaines « subtilités » qui se présentent à l'analyste au cours de la procédure de simulation.

Considérons une expérience dans laquelle chacune des n personnes testées doit lire un texte imprimé en caractères A et un texte de même complexité imprimé en caractères B. Dans les deux cas on note le temps t demandé par la lecture. Soient $t\ (i,\ A)$ et $t\ (i,\ B)$ le temps mis par la personne s pour lire les textes en caractères A et B respectivement. Cette situation peut être décrite par le modèle mathématique simple suivant:

$$\begin{aligned} t\ (i,\ A) &= m\ (i) + \tau_A + \xi_i, \\ i &= 1,\ 2,\ \ldots,\ n, \\ t\ (i,\ B) &= m\ (i) + \tau_B + \eta_i, \end{aligned} \tag{3.1}$$

où $m(i)$ est une variable aléatoire *) figurant la vitesse de lecture de la personne i indépendamment des caractères, τ_A et τ_B sont des constantes dépendant uniquement des caractères, ξ_i et η_i des erreurs aléatoires mutuellement indépendantes de moyenne 0 et de variance commune σ^2. Les quantités du second membre de l'équation (3.1) sont plus grandes que celle du premier. Ceci exprime qu'il est impossible d'estimer les principales caractéristiques numériques des quantités $m\ (i)$, τ_A, τ_B, ξ_i et η_i à l'aide des observations $\{t\ (i,\ A),\ t\ (i,\ B)\}_{i=1,\ldots,n}$. Bien plus, même si ξ_i et η_i ne figurent pas dans le modèle, c'est-à-dire dans le cas où le premier membre comprend $2n$ quantités et le second, $n + 2$, on ne peut trouver les principales caractéristiques numériques de m_i $(i = 1,\ 2,\ \ldots,\ n)$ sans

*) Les notions de variable aléatoire et de ses principales caractéristiques numériques (moyenne, variance, etc.) sont introduites au chapitre 5.

poser de conditions subsidiaires sur τ_A, τ_B. (Dans de tels cas, on dit parfois que le modèle est non mesurable par rapport aux données expérimentales.) Mais si le problème posé consiste seulement à comparer la vitesse moyenne de lecture des deux textes, alors la non-mesurabilité du modèle cesse d'être un obstacle. En effet, la variable aléatoire

$$t(i) = t(i, A) - t(i, B) = \tau_A - \tau_B + \xi_i - \eta_i \qquad (3.2)$$

a une moyenne strictement positive si les caractères B sont plus commodes à lire que les caractères A, et une moyenne strictement négative, dans le cas contraire. L'estimation de la différence $\tau_A - \tau_B$ basée sur les valeurs de $t(i)$ ne pose pas de problème. De façon analogue, si l'on avait à caractériser le degré de dispersion aléatoire de la vitesse de lecture de chaque personne (c'est-à-dire à estimer la *variance* $\mathbf{V}\, m(i)$; la définition et le calcul de $\mathbf{V}$ sont donnés aux chapitres 5 et 8), on aurait pu calculer la variance des variables aléatoires

$$\frac{1}{2}(t(i, A) + t(i, B)) = m(i) + (\tau_A + \tau_B)/2 + (\xi_i + \eta_i)/2 \qquad (3.3)$$

et en soustraire la quantité $\sigma^2/2$ qui définit la contribution des erreurs aléatoires dans le modèle (3.3). Ici on estime la variance de $m(i)$ sans déterminer $m(i)$ pour chaque i, mais en utilisant le fait que τ_A et τ_B sont des constantes.

Donc, les différentes variantes du modèle (3.1) *nous permettent de tenir compte*: de la différence des vitesses de lecture; de la différence entre la vitesse moyenne de lecture des caractères A et B; du caractère aléatoire du temps mis par la personne testée pour lire le texte. Mais ce modèle *néglige* la dépendance éventuelle de la différence $\tau_A - \tau_B$ par rapport à la vitesse de lecture $m(i)$ de i et à l'ordre dans lequel sont lus les textes. Par ailleurs, l'hypothèse de la constance de la variance σ^2 des erreurs aléatoires simplifie l'analyse. Il est évident que pour étudier plus en détail la durée, il aurait fallu un modèle plus compliqué qui reflétât les dépendances signalées. Mais, s'il s'agit seulement de comparer les vitesses moyennes de lecture des textes en caractères différents, le modèle (3.2) remplit bien cette fonction: il suggère une méthode assez efficace d'analyse des données en rejetant une autre approche (trop naïve) préconisant d'abord une moyennisation séparée des données par rapport à chaque caractère:

$$t(\cdot, A) = \sum_i t(i, A)/n\,; \quad t(\cdot, B) = \sum_i t(i, B)/n, \qquad (3.4)$$

et ensuite une comparaison des moyennes $t(\cdot, A)$ et $t(\cdot, B)$ obtenues soi-disant par deux séries indépendantes d'observations (cette comparaison peut être conduite, par exemple, à l'aide de la loi de Stu-

dent, cf. n° 11.2.8). La dernière méthode peut en pratique s'avérer peu efficace pour mettre en évidence la distinction existant entre les caractères A et B, car les observations $t(i, A)$ et $t(i, B)$ sont en fait étroitement liées à cause de la valeur commune $m(i)$.

Le modèle mathématique est à certains égards pour l'analyste ce que le laboratoire est pour le physicien. Il peut se livrer à des expériences sur le modèle mathématique, et si celui-ci reflète bien la réalité, alors *il applique* les résultats des expériences *au monde réel.*

Quand on dit qu'un modèle convient à la description de la réalité, on entend par là qu'il est possible de s'en servir comme point de départ pour choisir la meilleure méthode de traitement statistique des données initiales ainsi que pour la résolution de problèmes tels que la planification, les prédictions, la commande optimale de systèmes et de processus, l'estimation de l'efficacité du fonctionnement (ou la caractérisation de la qualité) d'un système complexe, les diagnostics (en médecine et technique), la normalisation.

3.2. Schéma logique général et principales étapes de modélisation mathématique

3.2.1. Principales étapes de modélisation. Au cours de la première étape il faut définir: les objectifs finals de la modélisation; les facteurs et les variables aux interdépendances desquels on s'intéresse; enfin, le rôle de ces facteurs et variables, plus exactement, lesquels d'entre eux peuvent être considérés dans le problème étudié comme variables *d'entrée* (c'est-à-dire susceptibles d'être régularisées totalement ou partiellement ou, à la rigueur, être facilement enregistrables et prévisibles; ces facteurs sont *explicatifs*), lesquels, comme variables *de sortie* (l'objectif final de l'analyse; ces facteurs sont en général difficilement enregistrables ou prévisibles et sont *à expliquer*). Si l'information statistique initiale n'a encore pas été recueillie, sa collecte est incorporée dans la première étape. Ainsi, dans l'exemple des caractères A et B, les diverses variantes des objectifs finals mettent en jeu des ensembles différents de facteurs analysés (type des caractères, la personnalité du testé) et de variables (d'une collection redondante de $3n + 2$ variables dans le modèle (3.1) on passe à 2 variables moyennisées dans (3.4)) en répartissant simultanément de manière différente les rôles entre eux.

Dans la deuxième étape on passe à la postulation, la formalisation mathématique et si possible à la vérification expérimentale des hypothèses de base concernant la nature et les propriétés physiques de l'événement analysé (étape d'élaboration de l'information initiale). Si les hypothèses adoptées ne peuvent pour une raison ou une autre faire l'objet d'une vérification expérimentale, il faut alors les asseoir par des raisonnements théoriques sur le mécanisme de l'événement étudié (elles peuvent, par exemple, être reconnues par les experts de

la discipline utilisatrice — économie, sociologie, technique, médecine, etc. — comme étant des lois objectives particulières). Ainsi, en construisant les modèles (3.1) à (3.4) nous sommes partis des conditions suivantes qui de prime abord semblent naturelles mais dont la validité est discutable: les composantes aléatoires ξ_i et η_i (les « erreurs d'enregistrement ») sont mutuellement indépendantes; les caractéristiques de la vitesse de lecture m (i) des testés ne dépendent pas de la nature des caractères; la différence $\tau_A - \tau_B$ ne dépend pas de l'ordre de lecture des textes, etc.

La troisième étape est l'*étape de modélisation proprement dite*, car elle comprend la déduction (d'après les hypothèses de base adoptées et confirmées en partie expérimentalement) de la forme *générale* des relations de fonctionnement qui lient les variables d'entrée et de sortie sélectionnées. Par détermination de la forme générale des relations de fonctionnement on entend la détermination, au niveau de cette étape, seulement de la structure du modèle, de sa représentation analytique symbolique, représentation dans laquelle conjointement aux valeurs numériques connues (figurées par les données statistiques initiales) participeront des variables dont la signification physique est connue mais pas les valeurs numériques (les paramètres inconnus du modèle qui feront l'objet d'une estimation statistique).

Dans l'exemple des caractères A et B la déduction des relations de fonctionnement (3.1) à (3.3) est triviale: elles résultent immédiatement des hypothèses et notations adoptées (en fait, ce n'est pas toujours le cas; cf., par exemple, la déduction des relations de fonctionnement décrivant le mécanisme des rapports distributifs dans la société dans [3]). Les premiers membres de ces relations contiennent des nombres connus (les données initiales), les seconds membres, les paramètres inconnus du modèle qui doivent être estimés statistiquement.

La quatrième étape ou étape d'*analyse statistique du modèle* est consacrée à la résolution du problème du meilleur choix, c'est-à-dire à l'estimation statistique des paramètres inconnus figurant dans la représentation analytique du modèle, et à l'étude des propriétés et de la précision des estimations obtenues. Ce problème se résout entièrement par les méthodes de traitement statistique des données.

Dans la cinquième étape ou étape de vérification du modèle, on applique diverses procédures de comparaison des conclusions et des estimations établies sur le modèle avec la réalité étudiée. Cette étape s'appelle aussi étape d'*analyse statistique de l'adéquation du modèle*.

La sixième étape est facultative en ce sens qu'elle dépend des résultats de l'étape précédente. Elle consiste en la planification et la réalisation de recherches destinées à améliorer le modèle et, en particulier, à approfondir la deuxième étape qui est *capitale* à certains égards.

3.2.2. Modélisation du mécanisme d'un événement au lieu d'une photographie statistique formelle. Arrêtons-nous plus en détail *sur l'aspect capital de la deuxième étape.* On affirme que l'adéquation et partant l'efficacité du modèle dépendent essentiellement de la profondeur et de la qualité de l'analyse de l'*essence réelle* de l'événement étudié pendant la collecte de l'information initiale (c'est-à-dire dans le cadre de la deuxième étape). En d'autres termes, lors de la modélisation probabilisto-statistique et, en particulier, lors de la collecte de l'information initiale sur la nature physique du mécanisme de transformation des variables d'entrée en variables de sortie, *une partie de ce mécanisme échappe à l'analyste* (cette partie est appelée « boîte noire » dans le langage des cybernéticiens). Plus l'analyste connaîtra à fond les mécanismes de fonctionnement du système étudié, plus il réduira la place occupée par la « boîte noire » dans le schéma logique général et plus le modèle construit sera opérationnel et précis. Une modélisation basée entièrement sur la logique de la « boîte noire » ne permet à l'analyste d'obtenir qu'une sorte de *photographie statistique instantanée* de l'événement étudié, qui en général est inapte par exemple aux prévisions. Au contraire, une modélisation bénéficiant d'une analyse *ex professo* de la nature de l'événement étudié facilite grandement la justification théorique de la forme générale du modèle construit et permet donc de l'appliquer largement au calcul des prévisions. Expliquons cette situation sur un exemple du § 6.1.

Supposons que l'on se fixe pour objectif la description laconique (*paramétrée à l'aide d'un modèle*) de la densité d'une variable aléatoire (représentant le salaire mensuel d'un employé pris au hasard dans une population générale) d'après les données initiales concernant un échantillon aléatoire d'employés $x_1, x_2, \ldots, x_{750}$ de taille $n = 750$. Si l'on néglige les lois économiques qui régissent la loi de répartition cherchée, c'est-à-dire si l'on s'en tient à la solution *formelle* du meilleur instantané statistique, on doit garder en réserve une classe assez vaste de densités modèles (par exemple, la classe des courbes de Pearson [76] comme ceci a été fait, en particulier, dans [28]) et étudier successivement ces modèles (en soumettant simultanément les paramètres de leurs relations de fonctionnement à une estimation statistique par les méthodes décrites au chapitre 8) pour dégager la densité qui approche le mieux selon un certain critère (par exemple, le critère du « khi-deux » de Pearson, cf. § 11.1) le comportement de la densité empirique (cf. représentation de l'histogramme correspondant sur la figure 6.1, page 139). Sur cette voie on peut, moyennant une extension de la classe des densités modèles hypothétiques, atteindre une très haute précision d'approximation, voire même la répétition, par la densité modèle, des creux inattendus de l'histogramme du genre de ceux observés sur les 14e et 15e intervalles de groupement de la figure 6.1. Mais en procédant ainsi, on obtient un résultat qui

n'est bon qu'en apparence, ce dont on s'assure sans peine en appliquant la loi modèle mise en évidence à la description de la densité empirique construite à l'aide d'un *autre échantillon parent*. Dans la grande majorité des cas, la densité modèle antérieure est inapte à décrire les lois de répartition observées dans un autre échantillon. Pour cet échantillon, il faut donc construire un *autre modèle*, ce qui vide la modélisation de sa substance, puisque la raison d'être d'un modèle est de généraliser les lois dégagées d'un échantillon à la population générale tout entière (tel est le cas dans les problèmes de planification, de prédiction, de diagnostic).

Une démarche opposée consiste à procéder préalablement à une analyse minutieuse et compétente des lois locales régissant la répartition du salaire mensuel. Ces lois (le caractère multiplicatif de la réduction du travail, le principe de rémunération en fonction du travail accompli, la constante du coefficient de variation du salaire lorsqu'on passe d'une catégorie d'employés à une autre, etc., cf. [3]) permettent dès l'étape suivante (la troisième) de la modélisation de justifier *théoriquement* (c'est-à-dire sans recourir à la densité empirique) le choix de la classe de modèles dans laquelle on cherchera la densité modèle. Dans l'exemple considéré, cette classe est la classe des distributions log-normales (cf. n° 6.1.6). On passe ensuite à l'estimation statistique des paramètres qui ont servi à décrire les lois de cette classe, c'est-à-dire à la quatrième étape.

Le modèle ainsi acquis approche en général plus mal (selon des critères formels) que le précédent la densité empirique construite *d'après l'échantillon donné*. Mais contrairement au modèle obtenu à la suite d'un ajustage statistique formel des données expérimentales sur une courbe théorique, il est *invariant par tout changement d'échantillon*, autrement dit, il décrit indifféremment bien le caractère de la distribution observée pour divers échantillons parents. Que dire d'une simulation basée sur des hypothèses plus ou moins crédibles (éventuellement, confirmées par l'expérience) sur la nature de l'événement étudié, qui donnerait des résultats s'accordant mal avec la réalité? La cause de ce désaccord est unique (sous réserve, bien sûr, que les troisième et quatrième étapes aient été conduites avec soin): violation de tout ou partie des hypothèses de base adoptées. Deux cas peuvent se présenter: si les hypothèses de base sont reconnues par les experts comme étant des lois objectives *gouvernant le fonctionnement* du mécanisme de l'événement étudié, alors il faut localiser et éliminer les causes de la violation de ces lois; si, au contraire, les hypothèses de base sont le résultat de la *simplification forcée* d'un mécanisme mal appréhendé, il faut améliorer ces hypothèses, ce qui nécessitera naturellement une modification du modèle. Dans l'exemple considéré, l'interprétation du désaccord provisoire entre le modèle et la réalité est du premier type. Les comparaisons des données fournies par les modèles et des données expérimentales sur la distri-

bution du salaire sur plusieurs années (1956-1972, cf. [3]) ont nettement mis en évidence une période de profond désaccord (1960-1963). Mais à mesure que l'on remonte dans le temps, on constate une tendance manifeste au rapprochement de ces données. Une analyse plus poussée a montré que ce désaccord faisait suite à une ingérence autoritaire dans le régime salarial, une ingérence qui, comme l'a montré l'avenir, a été contestée par de nombreuses lois économiques objectives. Et le fait que par la suite on assiste à un rapprochement des données réelles et des données modèles est significatif d'une chose, que ces lois économiques objectives ont pesé progressivement sur le caractère de la distribution, ont commencé à jouer davantage en s'érigeant en lois juridiques.

Les exemples de modélisation non formelle (mettant en évidence le mécanisme de l'événement étudié) ne sont malheureusement pas très nombreux (cf. [3], [80], etc.).

3.3. Notion de simulation statistique

L'étude d'un modèle probabiliste peut être conduite de deux façons : par les méthodes mathématiques de la théorie des probabilités et de statistique mathématique et par la reproduction du fonctionnement de ce modèle sur ordinateur. Le second procédé s'appelle *simulation statistique*. Pour certains auteurs ([30], [51]), ce terme recouvre une idée plus large : il comprend en plus la réduction de problèmes traditionnels de calcul numérique à des modèles probabilistes par simulation sur ordinateur. Bien que cette interprétation augure d'importants et intéressants développements, nous nous bornerons à la première définition, plus étroite, de la simulation statistique.

La simulation statistique est un puissant outil de manipulation des modèles probabilistes à toutes les étapes de la recherche. En théorie des files d'attente, par exemple, c'est la principale méthode de résolution des systèmes complexes [30]. En statistique classique, c'est l'une des méthodes d'étude de la stabilité des estimations par rapport aux fluctuations des hypothèses de base, une méthode utilisée seule ou conjointement à des méthodes analytiques asymptotiques. Si le modèle de l'événement étudié est relativement compliqué, la simulation statistique peut aider à déterminer la taille des échantillons principal et de contrôle. Elle est également largement appliquée pour étudier les limites tolérables des écarts entre les hypothèses adoptés pendant la construction d'un modèle et la réalité. Les perspectives offertes par la simulation statistique dans l'analyse des modèles probabilistes sont si riches que d'ores et déjà elles rendent indispensable la justification de l'appareil analytique traditionnellement utilisé pour leur construction et l'étude de leurs propriétés. Au § 6.3 figure une brève description d'une technique de simulation statistique d'observations suivant une loi de probabilité donnée.

3.4. Objections contre les modèles mathématiques

A l'heure actuelle les modèles mathématiques ont reçu leurs lettres de noblesse dans de nombreux domaines des sciences et des techniques. Mais certains esprits, notamment des médecins, des sociologues, des économistes, affirment encore que les *phénomènes étudiés sont par trop complexes pour être adéquatement exprimés par des méthodes mathématiques.* Sans nier les traits spécifiques et la complexité des phénomènes étudiés par ces corps de spécialistes, il faut affirmer avec conviction qu'il n'y a pas de substitut au langage des modèles mathématiques et à la simplification et la schématisation de la réalité qu'ils nécessitent. Mais ce qui tient les rênes courtes aux chercheurs est moins dans les performances de l'appareil mathématique que dans les capacités de l'homme à maîtriser les situations complexes. En effet, à quoi bon concevoir des modèles mathématiques sophistiqués si nous ne pouvons en « digérer » les fruits Une solution qui permet d'éviter les extrêmes simplifications des phénomènes étudiés consiste à envisager plusieurs modèles à la fois, chacun de ces modèles pouvant être regardé comme un cas particulier d'un modèle plus général nécessaire pour décrire la réalité de façon adéquate.

La *deuxième objection* contre l'implantation des modèles probabilistes dans les sciences étudiant le comportement de l'homme est le fait d'une interprétation erronée du rôle des probabilités dans les modèles. Une partie des savants inclinent encore à penser que considérer l'homme comme représentant d'une population générale, comme un individu dont les réactions sont décrites par des lois probabilistes, reviendrait à le priver de sa personnalité et de sa liberté de choix. De fait, c'est justement la personnalité de l'homme et l'imprévision de son choix qui plaident pour une description de son comportement en termes probabilistes. L'usage des probabilités pour décrire un modèle ne veut nullement dire que chaque individu définit son propre comportement à l'aide d'un générateur de nombres aléatoires. Cela signifie tout simplement que le comportement d'un groupe donne l'impression que les individus utilisent effectivement des générateurs de nombres aléatoires. La fonction des notions probabilistes est de décrire les variations des réactions. Rien de commun donc avec la restriction des libertés de choix, ou la contrainte de la personnalité des sujets étudiés. La seule alternative à l'approche probabiliste est l'approche déterministe, or c'est justement elle qui prive l'homme de son individualité. Les modèles déterministes décrivant de façon simplifiée le comportement d'un individu sont très « rentables » dans l'analyse des événements de masse. A titre d'exemple, citons les recherches effectuées pour construire un modèle de propagation d'une épidémie de grippe [17].

L'implantation massive des ordinateurs est à l'origine de la *troisième objection, la plus perfide* semble-t-il dans la mesure où elle vise non pas les modèles proprement dits mais l'adéquation de leur étude par les méthodes mathématiques. Est-il besoin d'accomplir un travail de bénédictin, d'étudier des modèles spéciaux en théorie des files d'attente, alors que la modélisation statistique répond à toute question plus vite et à moindre frais? S'il s'agit d'obtenir simplement une réponse à un cas particulier, la modélisation statistique est la meilleure méthode. Mais, si l'objectif est la solution générale et la connaissance à fond du phénomène étudié, alors la modélisation statistique manque de souffle. Une conclusion résumée par une formule claire et simple justifie tous les efforts déployés sur le plan théorique même lorsqu'on est contraint à des simplifications pendant la construction et l'étude du modèle.

3.5. Types de modèles mathématiques les plus souvent utilisés en statistique appliquée

3.5.1. Modèles des lois de probabilité de variables aléatoires. La notion de loi de probabilité d'une variable aléatoire est abordée au § 5.4. Par *variable aléatoire* on entend une variable recouvrant une certaine idée physique, dont les valeurs sont affectées d'une dispersion incontrôlable lors de la répétition de l'expérience étudiée ou des observations (cf. § 5.1). Connaître la loi de probabilité d'une variable aléatoire, c'est pouvoir associer à n'importe laquelle de ses valeurs éventuelles (ou à n'importe quel domaine de ses valeurs éventuelles) la probabilité d'apparition de cette valeur (ou la probabilité qu'elle prenne une valeur de ce domaine).

L'analyse statistique commence en principe par l'analyse de la loi de probabilité de la variable aléatoire considérée, par la construction d'un modèle de cette loi. Au chapitre 6 sont décrites les lois de probabilité les plus couramment utilisées dans les recherches statistiques.

3.5.2. Modèles probabilistes linéaires. Les *modèles de régression linéaire* émergent du lot de modèles décrivant la dépendance des variables aléatoires. Ils sont généralement de la forme

$$Y = X'\theta + \varepsilon, \tag{3.5}$$

où $Y = (y_1, \ldots, y_n)'$ est le vecteur n-dimensionnel des observations ; X une matrice du plan connue de dimension $p \times n$; $\Theta = (\theta_1, \ldots, \theta_p)'$ un vecteur des paramètres p-dimensionnel inconnu ; ε un vecteur colonne aléatoire n-dimensionnel vérifiant la condition

$$\mathrm{E}(\varepsilon) = 0, \ \mathrm{E}(\varepsilon\varepsilon') = \sigma^2 \mathrm{I}, \tag{3.6}$$

où σ^2 est un paramètre scalaire inconnu, E le symbole de l'espérance mathématique (cf. n° 5.6.1). Souvent y_i, $i = 1, \ldots, n$, sont interprétés comme les observations de la *variable prédictante* (la *réponse*) au point $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})'$ de l'espace des observations.

Montrons tout d'abord que le modèle des caractères A et B étudié ci-dessus (cf. § 3.1) peut être traité comme un cas particulier d'un modèle linéaire général. Pour cela posons $\theta_i = m(i)$, $i = 1, 2, \ldots$ $\ldots, n$; $\theta_{n+1} = \tau_A$; $\theta_{n+2} = \tau_B$; $y_i = t(i, A)$, $y_{n+i} = t(i, B)$. Les équations (3.1) peuvent être mises sous la forme (3.5) à l'aide d'une matrice X de dimension $(n + 2) \times 2n$ telle que

$$X' = \begin{vmatrix} 1 & 0 & 0 & \ldots & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & \ldots & 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 0 & 1 & 0 & 1 \end{vmatrix}.$$

L'hypothèse nulle dans cette paramétrisation consiste à vérifier l'égalité $\theta_{n+1} = \theta_{n+2}$.

Comme autres cas particuliers du modèle (3.5), (3.6) signalons:

a) le *modèle de régression linéaire du premier ordre* qui met en jeu un seul prédicteur quantitatif (un facteur). Lorsque ce prédicteur est égal à x, la variable prédictante (ou réponse) est égale à

$$y = \theta_0 + \theta_1 x + \varepsilon; \tag{3.7}$$

b) le *modèle d'analyse de variance monofactorielle* qui fait intervenir un prédicteur (qualitatif) à l modalités et n_l observations indépendantes de chaque modalité:

$$y_{ik} = \theta_0 + \theta_i + \varepsilon_{ik}, \quad i = 1, \ldots, l; \; k = 1, \ldots, n_l. \tag{3.8}$$

Pour que ce modèle admette une solution, il faut conjecturer en plus que $\sum_{i=1}^{l} \theta_i = 0$. On s'intéresse le plus souvent à la question de savoir si tous les θ_i, $i = 1, \ldots, n$, sont nuls;

c) le *modèle d'analyse bifactorielle*. Le nom lui-même indique qu'on a affaire à deux prédicteurs (qualitatifs). La réponse à la i-ième composante du premier prédicteur et à la j-ième du second est de la

forme

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, \ldots, l; \quad j = 1, \ldots, m,$$

où les prédicteurs doivent satisfaire les conditions subsidiaires $\sum \alpha_i = 0$, $\sum \beta_j = 0$ et ξ_{ij} sont des erreurs indépendantes de même loi de probabilité. Les hypothèses les plus souvent vérifiées sont : $\alpha_1 = \ldots = \alpha_l = 0$, $\beta_1 = \ldots = \beta_m = 0$.

Les modèles linéaires sont bien étudiés, cf. par exemple, [114], [117].

3.5.3. Généralisation des modèles linéaires. Décrivons les deux desserrements des hypothèses adoptées dans le modèle (3.5), (3.6), les plus couramment utilisés. Ces desserrements permettent de rendre le modèle plus conforme à la réalité.

1. *La fonction décrivant la dépendance de la réponse y par rapport aux variables X n'est pas linéaire en les paramètres* Θ. Cela exprime que la variable y observée est une fonction des variables observées $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ et des paramètres inconnus Θ, soit

$$Y = \Psi(\Theta, X) + \varepsilon. \tag{3.9}$$

Le vecteur des erreurs aléatoires est soumis à des contraintes de la forme (3.6). Ces modèles font recette dans l'étude des phénomènes économiques, physiques, chimiques, etc., dont le mécanisme de fonctionnement n'est pas linéaire. L'analyse des modèles (3.9) passe généralement par la linéarisation de ψ au voisinage des valeurs prévisibles de Θ et l'étude du modèle linéaire ainsi obtenu par la méthode des moindres carrés (cf. § 8.6).

2. *La dispersion de l'erreur dépend des paramètres à estimer.* Ce desserrement des hypothèses imposées au modèle est dicté comme plus haut par des considérations propres à la discipline utilisatrice. Pour estimer les paramètres, on se sert le plus souvent de la méthode du *maximum de vraisemblance* (cf. n° 8.6.1). Le système des équations du maximum de vraisemblance se résout itérativement par des linéarisations successives. A chaque itération le poids des observations est considéré comme un nombre donné qui est défini par la valeur des paramètres à l'itération précédente.

Dans les modèles linéaires, la matrice X est supposée connue et fixe. Mais dans certaines recherches statistiques, telles les mesures des trajectoires de particules élémentaires en physique, l'enregistrement des composantes de réactions chimiques complexes, etc., les valeurs des variables X ne peuvent être strictement fixées et doivent être considérées comme les moyennes inconnues de variables *aléatoires* obéissant à une certaine loi de probabilité d'une mesure élémentaire à l'autre. Dans ces conditions, la séparation du rôle des variables indépendantes x et des variables dépendantes y du modèle devient inadéquate. Les deux suites (x_i, y_j) sont considérées comme

aléatoires, elles *fusionnent* en quelque sorte du point de vue de leur interprétation. Les modèles correspondants sont appelés, suivant Frisch, modèles *confluents* et les méthodes qui les étudient, méthodes de l'*analyse confluente*.

Voici un exemple élémentaire de modèle confluent: on observe des couples $(\tilde{x}_i, \tilde{y}_i)$ tels que

$$\tilde{x}_i = x_i + \varepsilon_i; \ \tilde{y}_i = y_i + \delta_i; \ y_i = a + bx_i, \tag{3.10}$$

où x_i, y_i sont les valeurs vraies inconnues des variables; a et b les paramètres inconnus reliant les variables x et y; ε_i, δ_i des erreurs aléatoires indépendantes normales, de moyenne nulle et de variances connues.

Parfois l'analyse confluente est appelée *analyse des relations structurales* à cause de (3.10).

3.5.4. Modèles géométriques. Dans la confection sur mesure d'une robe il faut prendre plus d'une quinzaine de mesures de la cliente. Mais ces mesures ne sont pas toutes de la même importance pour la qualité de l'article, ce qui permet de se limiter à quatre d'entre elles dans la confection en série: la taille, le tour de poitrine, le tour de taille, le tour de hanches. Essayons d'interpréter ce fait géométriquement. A cet effet, on représentera le vecteur des mesures d'une femme par un point d'un espace R^p $(p > 10)$ à p dimensions et on considérera l'ensemble des points figurant une importante population de femmes. Ces points ne sont pas répartis de façon uniforme dans l'espace R^p et sont groupés autour d'une surface à quatre dimensions dont chaque point peut être caractérisé par quatre des dix mesures de base. Les modèles géométriques servent précisément à décrire des situations dans lesquelles il est possible de dégager dans l'espace des mesures de base une surface de moindre dimension autour de laquelle sont groupés les points initiaux. On peut en dire autant des modèles linéaires. Mais la classe des modèles géométriques est plus vaste et l'on a élaboré des méthodes spéciales pour l'étudier (cf. § 10.5). Par ailleurs, le premier examen des données dans le but de dégager leurs principales tendances de variation précède généralement la construction de modèles paramétriques, y compris de modèles linéaires.

Un modèle géométrique élémentaire est de la forme

$$X = A + \Xi, \quad A \in V_r, \ \mathbf{E}\Xi = 0, \tag{3.11}$$

où X est le vecteur p-dimensionnel des observations initiales, A un vecteur p-dimensionnel aléatoire dont on sait seulement qu'il appartient à une surface V_r de dimension intérieure $r < p$ et Ξ un vecteur aléatoire p-dimensionnel décrivant l'écart de X par rapport à V_r. Souvent, on admet encore que $\mathbf{E}(\Xi', \Xi)$ est petite en regard de l'éten-

due de A sur V_r. Le modèle (3.11) présente de nombreux traits spécifiques :

1) on n'exige pas que Ξ varie uniquement dans une direction perpendiculaire à V_r ;

2) aucune condition n'est posée quant à la forme de V_r. On aurait pu admettre, par exemple, que V_r est une hypersurface à r dimensions ;

3) la disposition des vecteurs A sur V_r n'est soumise à aucune condition. On peut, par exemple, supposer que les vecteurs A sont concentrés sur plusieurs ellipsoïdes isolés à r dimensions, etc.

Comme cas particulier du modèle (3.11), on peut considérer le modèle (3.10). Dans ce cas, $p = 2$, $r = 1$, $X = (\tilde{x}, \tilde{y})'$, V_r est confondue avec la droite $y = a + bx$, et Ξ est le vecteur $(\varepsilon, \delta)'$.

3.5.5. **Modèles markoviens.** En sociologie, économie, médecine, démographie, on utilise largement les modèles décrivant la dynamique d'indices économiques et sociaux au moyen des probabilités de passage d'une structure du système réel étudié à une autre. On exploite ici l'appareil des chaînes de Markov discrètes et continues avec une paramétrisation linéaire et non linéaire des probabilités de passage [18], [126]. Les modèles envisagés sont très complexes et leur analyse implique souvent une approche individuelle et beaucoup de doigté dans la manipulation des principes fondamentaux de la statistique mathématique.

Conclusions

1. Les modèles probabilisto-statistiques sont un puissant instrument de description qualitative des relations liant les phénomènes et faits observés, d'étude des propriétés des systèmes envisagés, de choix d'un appareil statistique adéquat pour le traitement des données et l'organisation de la collecte des données.

2. Les modèles probabilisto-statistiques sont étudiés aussi bien avec l'arsenal traditionnel de la statistique mathématique qu'avec la simulation statistique qui consiste à simuler sur ordinateur le fonctionnement du modèle.

3. Tout modèle mathématique est une représentation simplifiée de la réalité et tout l'art du chercheur est de conjuguer la paramétrisation la plus laconique du modèle à une description assez adéquate de la réalité, en d'autres termes, il doit « compresser » au maximum la réalité dans une forme mathématique simple.

4. La procédure de modélisation peut être conventionnellement partagée en *six étapes principales* : première étape — définition des objectifs de la modélisation, des variables du modèle et de leur rôle ; deuxième étape — analyse de la nature physique du phénomène étudié, collecte et formalisation de l'information initiale ; troisième

étape — modélisation proprement dite (détermination de la forme générale du modèle); quatrième étape — analyse statistique du modèle (estimation des paramètres participant à la description du modèle); cinquième étape — vérification du modèle; sixième étape (facultative) — précision du modèle et notamment retour à la deuxième étape.

5. Une condition *sine qua non* du bon fonctionnement d'un modèle est la réalisation de la deuxième étape de la modélisation, c'est-à-dire la réalisation d'une analyse minutieuse de la nature du phénomène étudié afin de recueillir une information initiale de bon aloi et d'en tirer le meilleur parti pour la déduction (ou le choix) de la forme générale du modèle cherché. La contre-approche forcée (mais indésirable) est la logique de la « boîte noire », c'est-à-dire une approximation purement formelle des données réelles.

6. Les modèles mathématiques les plus couramment utilisés dans les applications statistiques sont: les modèles des lois de probabilité; les modèles linéaires décrivant le caractère et la structure des relations entre les variables analysées (notamment les modèles de régression, d'analyse de variance, les modèles d'analyse factorielle et de séries temporelles, cf. par exemple [114]); les modèles markoviens décrivant les lois des passages aléatoires d'objets d'un état à un autre; les modèles géométriques qui permettent de visualiser les données multidimensionnelles initiales (cf. § 10.5).

7. Ces deux dernières décennies, grâce aux performances des ordinateurs, l'analyste ne s'en tient plus qu'aux modèles facilement résolubles, il peut envisager diverses généralisations de modèles linéaires reflétant la réalité de façon plus adéquate. Dans l'analyse mathématique des modèles, on constate une tendance à éviter toute relation linéaire entre l'espérance mathématique du vecteur d'observations Y et les paramètres Θ du modèle; à faire dépendre la variance de la réponse des valeurs des paramètres; à renoncer à une matrice $\mathbf{X}$ dont les éléments sont supposés connus à une erreur aléatoire près. L'estimation des paramètres du modèle passe par la résolution des équations du maximum de vraisemblance.

DEUXIÈME PARTIE

FONDEMENTS MATHÉMATIQUES DE LA THÉORIE DES PROBABILITÉS

CHAPITRE 4

OPÉRATIONS SUR LES ÉVÉNEMENTS ALÉATOIRES ET LES PROBABILITÉS DE LEUR RÉALISATION

4.1. Espace probabilisé discret

4.1.1. Expérience aléatoire. Quand on élabore une théorie, on doit tout d'abord s'entendre sur les *définitions* de certaines notions et sur l'adoption d'hypothèses sans démonstration (*axiomes*). La notion liminaire de la théorie des probabilités est la notion d'*expérience aléatoire*. On appelle expérience aléatoire l'observation réalisée sur une unité de la population analysée dans les conditions d'ensemble statistique. La teneur de cette notion conditionne l'appareil mathématique utilisé pour la résolution du problème posé. *Dans le cadre d'un même problème, la notion d'expérience aléatoire s'interprète différemment selon les objectifs finals.* Ainsi, quand on contrôle des pièces d'après le principe binaire (on entend par là que la pièce testée est jugée ou bonne ou mauvaise), cette notion admet au moins trois interprétations : a) contrôle d'une seule pièce ; b) contrôle d'un lot (d'un échantillon) de N pièces ; c) contrôle de deux lots (échantillons) respectivement de N_1 et N_2 pièces (échantillon double). Il est évident que la première variante donne lieu à deux issues seulement (bonne ou mauvaise) ; la deuxième à $N + 1$ issues possibles (0, ou 1, ou 2, . . ., ou N pièces défectueuses) si l'on s'intéresse au nombre total de pièces défectueuses et non à l'ordre de leur mise en évidence ; la troisième, dans la même approche, à $(N_1 + 1)(N_2 + 1)$ issues possibles (le résultat de chaque expérience aléatoire est donné dans ce cas par un couple de nombres (ξ, η), où ξ et η sont les nombres de pièces défectueuses trouvées dans le premier et dans le second lot).

De façon exactement analogue, quand on analyse les résultats du lancement d'un dé, on peut comprendre par expérience aléatoire un jet (qui donnera lieu à six issues) ou une série d'un nombre donné de jets comme nous aurions dû le faire dans le cadre d'une formalisation rigoureuse de l'exemple du § 2.2 (on établit sans peine que le nombre total d'issues possibles est égal à $6^4 = 1296$).

Il n'est pas toujours possible de calculer le nombre d'issues d'une expérience aléatoire. Toute expérience aléatoire impliquant la fixation d'un paramètre quelconque de l'objet étudié, figurant une

quantité physique *continue* (température, pression, temps, poids, dimensions, etc.) possède un ensemble d'issues *) qui a la puissance du continu. Au paragraphe suivant, on verra comment la théorie des probabilités a été généralisée à ce cas.

4.1.2. Evénements aléatoires et opérations sur eux. On a déjà vu qu'à chaque expérience aléatoire était rattaché un ensemble d'issues possibles. Chaque issue s'appelle *événement* (ou *issue*) *élémentaire*, et l'ensemble de toutes les issues, *espace des événements* (*issues*) *élémentaires* ou *espace fondamental*. Ainsi, chaque expérience aléatoire donne obligatoirement lieu à un événement élémentaire et à un seul.

Etudions pour l'instant le *cas discret* seulement, c'est-à-dire les situations où les événements élémentaires peuvent être numérotés par les nombres 1, 2, . . ., n, . . . En d'autres termes, l'espace des événements élémentaires (que nous désignerons par Ω) n'est composé que d'un nombre fini ou dénombrable d'événements élémentaires (appelés ω_1, ω_2, . . ., ω_n, . . .). On note ceci

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$$

ou encore

$$\Omega = \{\omega_i\}, \quad i = 1, 2, \ldots \tag{4.1}$$

Voyons quelques exemples d'expériences aléatoires et d'espaces des événements élémentaires correspondants.

Exemple 4.1. Lancement d'une pièce de monnaie:

$$\Omega = \{\omega_1 = \text{face};\ \omega_2 = \text{pile}\}.$$

Exemple 4.2. Jet d'un dé:

$$\Omega = \{\omega_1 = 1;\ \omega_2 = 2;\ \omega_3 = 3;\ \omega_4 = 4;\ \omega_5 = 5;\ \omega_6 = 6\}.$$

Exemple 4.3. Série de quatre jets de dé:

$$\Omega = \{\omega_1 = (1, 1, 1, 1);\quad \omega_2 = (1, 1, 1, 2);\ \ldots,\ \omega_{1296} = (6, 6, 6, 6)\}.$$

Exemple 4.4. Test (d'après le principe binaire) d'un article choisi au hasard dans un lot produit en série:

$$\Omega = \{\omega_1 = \text{bon};\ \omega_2 = \text{mauvais}\}.$$

Exemple 4.5. Test (d'après le principe binaire) de N articles choisis au hasard dans un lot produit en série:

$$\Omega = \{\omega_1 = 0;\ \omega_2 = 1;\ \ldots;\ \omega_{N+1} = N\},$$

où ω est le nombre d'articles défectueux trouvés.

*) On sait qu'un ensemble peut être *fini* (si ses éléments sont en nombre fini), *dénombrable* (si ses éléments peuvent être numérotés par les nombres 1, 2, . . ., n, . . ., tels l'ensemble des entiers, l'ensemble des rationnels, etc.) ou avoir la *puissance du continu* (tel l'ensemble de tous les points de la droite numérique, du plan, etc.).

Exemple 4.6. Test (d'après le principe binaire) de deux échantillons de N_1 et N_2 articles choisis au hasard dans une production en série :

$$\Omega = \{\omega_1 = (0, 0);\ \omega_2 = (0, 1);\ \ldots;\ \omega_{(N_1+1)(N_2+1)} = (N_1, N_2)\}.$$

Exemple 4.7. Détermination du nombre ω de pannes de machines-outils servies par une équipe prise au hasard :

$$\Omega = \{\omega_1 = 0;\ \omega_2 = 1;\ \ldots;\ \omega_n = n - 1;\ \omega_{n+1} = n;\ \ldots\}.$$

Outre les événements élémentaires, nous aurons souvent à considérer des événements composés.

On dit qu'un événement C est *composé* (ou *décomposable*) si l'on peut indiquer au moins deux événements élémentaires ω_{i_1} et ω_{i_2} dont la réalisation de l'un d'eux entraîne celle de l'événement C. On exprime ceci en disant que « l'événement C est composé des événements élémentaires ω_{i_1} et ω_{i_2} » et on note $C = \{\omega_{i_1}, \omega_{i_2}\}$.

Dans le cadre de cette terminologie, on appellera *événement aléatoire A tout sous-ensemble* $\{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}, \ldots\}$ *de l'espace fondamental* (4.1), soit

$$A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}, \ldots\}.$$

Ceci s'explicite comme suit : la réalisation de n'importe quel événement élémentaire $\omega_{i_1}, \ldots, \omega_{i_k}$ « composant A » entraîne celle de l'événement A.

Précisons cette terminologie sur les exemples envisagés ci-dessus.

Dans l'exemple 4.2 les événements $A_1 = \{$le numéro sortant est pair$\}$ et $A_2 = \{$le numéro sortant est inférieur à 3$\}$ s'écrivent respectivement $A_1 = \{\omega_2, \omega_4, \omega_6\}$ et $A_2 = \{\omega_1, \omega_2, \omega_3\}$.

Dans l'exemple 4.3, l'événement $A = \{$le six sortira une fois au moins dans une série de quatre jets$\}$ sera composé de tous les vecteurs ω_i à quatre dimensions dont l'une au moins des composantes est égale à 6 (ces vecteurs sont au nombre de 671).

Dans l'exemple 4.5, l'événement $A = \{$le nombre d'articles défectueux est inférieur à 4$\}$ s'écrit visiblement $A = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$.

Dans l'exemple 4.7, l'événement $A = \{$le nombre de pannes de machines-outils servies par une équipe est strictement inférieur à 3$\}$ est de la forme $A = \{\omega_1, \omega_2, \omega_3\}$.

De la définition de la théorie des probabilités (cf. § 2.1) il s'ensuit qu'il faut en premier lieu convenir d'une certaine structure des relations entre les événements aléatoires ou, en d'autres termes, de *certaines opérations sur les événements* et d'une terminologie adéquate.

Somme (union) d'événements. La somme des événements A_1, $A_2, \ldots, A_k$ est l'événement A (noté $A = A_1 + A_2 + \ldots + A_k$)

qui consiste en la réalisation de l'un au moins des événements $A_1, A_2, \ldots, A_k$. Dans le langage des événements élémentaires, la somme des événements $A_1, A_2, \ldots, A_k$ se définit donc comme l'événement A composé de *tous les événements élémentaires* constituant les événements $A_1, A_2, \ldots, A_k$. Si, par exemple, on s'intéresse à la somme A des événements $A_1 = \{$le nombre de pièces défectueuses est inférieur à $4\} = \{\omega_1, \omega_2, \ldots, \omega_5\}$ (cf. exemple 4.5) et $A_2 = \{$le nombre de pièces défectueuses est compris entre 2 et 6 inclus$\} = \{\omega_3, \omega_4, \omega_5, \omega_6, \omega_7\}$, alors de toute évidence $A = A_1 + A_2 = \{\omega_1, \omega_2, \ldots, \omega_5, \omega_6, \omega_7\}$, car la réalisation de l'événement A sera signifiée par celle de l'un quelconque des événements $\omega_1, \ldots, \omega_7$.

Produit (intersection) d'événements. Le produit des événements $A_1, A_2, \ldots, A_k$ est l'événement A (noté $A = A_1 \cdot A_2 \cdot \ldots \cdot A_k$) qui consiste en la réalisation simultanée de tous les événements $A_1, A_2, \ldots, A_k$. Dans le langage des événements élémentaires, le produit des événements $A_1, A_2, \ldots, A_k$ se définit donc comme l'événement A qui est composé des seuls événements élémentaires *communs à tous les événements* $A_1, A_2, \ldots, A_k$. Ainsi (cf. exemple 4.5), le produit des événements $A_1 = \{\omega_1, \omega_2, \ldots, \omega_5\}$ et $A_2 = \{\omega_3, \omega_4, \ldots, \omega_7\}$ est visiblement l'événement $A = A_1 \cdot A_2 = \{\omega_3, \omega_4, \omega_5\}$, puisque la réalisation de chacun des événements élémentaires ω_3, ω_4 et ω_5 séparément signifie de toute évidence la réalisation simultanée des événements A_1 et A_2.

Différence d'événements. La différence de deux événements A_1 et A_2 est l'événement A consistant en la réalisation simultanée des deux faits suivants: l'événement A_1 s'est produit, l'événement A_2 ne s'est pas produit. Dans le langage des événements élémentaires, la différence $A = A_1 - A_2$ des événements élémentaires A_1 et A_2 se définit donc comme l'événement composé de tous les événements élémentaires figurant dans A_1 et pas dans A_2. Ainsi, la différence des événements $A_1 = \{\omega_1, \omega_2, \ldots, \omega_5\}$ et $A_2 = \{\omega_3, \omega_4, \ldots, \omega_7\}$ de l'exemple 4.5 sera visiblement l'événement $A = A_1 - A_2 = \{\omega_1, \omega_2\}$, c'est-à-dire l'événement qui consiste en ce que le nombre de pièces défectueuses est inférieur à 1.

Evénement contraire (complémentaire). L'événement $\bar{A}$ (lire non A) contraire de A consiste en la non-réalisation de A. Dans le langage des événements élémentaires, $\bar{A}$ se définit donc comme l'événement composé de tous les événements élémentaires ne figurant pas dans A. La notion de différence d'événements nous permet de noter ceci de la manière suivante: $\bar{A} = \Omega - A$. Ainsi, dans l'exemple 4.5, le complémentaire de l'événement $A_1 = \{\omega_1, \omega_2, \ldots, \omega_5\}$ est de toute évidence l'événement qui consiste en ce que le nombre N des articles défectueux est strictement supérieur à 4, c'est-à-dire que $\bar{A}_1 = \{\omega_6, \omega_7, \ldots, \omega_N\}$.

Evénement certain. C'est l'événement constitué de tous les événements élémentaires, soit $\Omega = \{\omega_i\}$, $i = 1, 2, \ldots$ (cf. (4.1)). Cet événement doit son nom au fait que l'espace des événements élémentaires est composé de *toutes* les issues possibles, autrement dit, l'expérience aléatoire analysée donne *nécessairement* lieu à un événement élémentaire ω_i, et par suite le fait que Ω a lieu est certain.

Evénement impossible (vide) $\varnothing$. C'est l'événement contraire de l'événement certain, c'est-à-dire que $\varnothing = \Omega - \Omega = \overline{\Omega}$. La dénomination de cet événement résulte immédiatement de sa définition : *cet événement ne contient aucun événement élémentaire* ω_i, donc, l'expérience réalisée est impossible.

Evénements incompatibles. Des événements $A_1, A_2, \ldots, A_k$ sont incompatibles s'ils ne peuvent se produire simultanément deux à deux. Dans le langage des événements élémentaires, cela signifie que parmi les événements $A_1, A_2, \ldots, A_k$ il n'existe aucun couple A_i, A_j ayant au moins un événement élémentaire en commun. Les notions de produit d'événements et d'événement impossible nous permettent de définir les événements incompatibles comme une suite d'événements $A_1, A_2, \ldots, A_k$ telle que $A_i \cdot A_j = \varnothing$ quels que soient A_i et A_j. Il est évident que deux événements contraires A et $\overline{A}$ sont incompatibles. C'est le cas aussi de tous les événements élémentaires (par définition).

Système complet d'événements. C'est un ensemble d'événements incompatibles $A_1, A_2, \ldots, A_k$ tel que

$$\begin{cases} A_1 + A_2 + \ldots + A_k = \Omega\,; \\ A_i \cdot A_j = \varnothing \end{cases}$$

pour tous les $i, j = 1, 2, \ldots, k$ et $i \neq j$.

Il est évident que l'ensemble fondamental peut être considéré comme un cas particulier d'un système complet d'événements. D'une façon générale, la partition de l'ensemble fondamental Ω en classes disjointes nous donne un système complet d'événements dans lequel chaque événement est défini par la classe correspondante. Revenons à l'exemple 4.5. Pour fixer les idées, posons $N = 100$, désignons par d le nombre de pièces défectueuses trouvées dans un échantillon de 100 pièces prélevées au hasard dans une production en série fonctionnant en régime stationnaire et définissons un système d'événements comme suit :

$$A_1 = \{d = 0\} = \{\omega_1\} \text{ un lot d'excellente qualité ;}$$

$$A_2 = \{1 \leqslant d \leqslant 5\} = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$

un lot présentant des défauts tolérables (le lot est acceptable) ;

$$A_3 = \{6 \leqslant d \leqslant 10\} = \{\omega_7, \omega_8, \ldots, \omega_{11}\}$$

un lot présentant un taux de pièces défectueuses supérieur à la norme admise (ce lot est accepté à un tarif bas);

$$A_4 = \{d \geqslant 11\} = \{\omega_{12}, \omega_{13}, \ldots, \omega_{101}\}$$

un lot mis au rebut.

Il est évident que les événements A_1, A_2, A_3 et A_4 forment un système complet.

4.1.3. Espace probabilisé. Probabilités et opérations sur elles. La donnée de l'ensemble fondamental ne suffit pas à elle seule à décrire entièrement le mécanisme de l'expérience aléatoire étudiée. Il est évident qu'en plus de toutes les issues de cette expérience nous devons connaître aussi la fréquence d'apparition des divers événements élémentaires. En effet, en revenant, pour fixer les idées, aux exemples 4.1 à 4.7, on conçoit sans peine que, dans le cadre de chacun des ensembles fondamentaux décrits, on peut considérer une infinité d'expériences différant fondamentalement par leur mécanisme. Dans les exemples 4.1 à 4.3, les mêmes événements élémentaires se produiront avec des fréquences différentes si l'on se sert de pièces de monnaie ou de dés différents (dés symétriques, dés à centre de gravité légèrement ou fortement décalé, etc.). Dans les exemples 4.4 à 4.7, la fréquence d'apparition des articles défectueux, le caractère de la pollution des lots testés par les articles défectueux et la fréquence d'apparition d'un certain nombre de pannes de machines-outils dépendront du niveau technologique de la branche étudiée: avec un même ensemble fondamental la fréquence d'apparition des « bonnes » issues élémentaires sera supérieure dans des processus dont le niveau technologique est plus élevé.

Pour bâtir (dans le cas discret) une théorie mathématique de l'expérience aléatoire (la théorie des probabilités) qui soit complète et cohérente, il faut en plus des notions déjà introduites d'*expérience aléatoire*, d'*issue élémentaire* et d'*événement aléatoire* se prévaloir d'un *axiome* postulant l'existence des probabilités des événements élémentaires (ces probabilités devant satisfaire une condition de normalisation) et de la *définition* de la probabilité de tout événement aléatoire.

Axiome. A chaque élément ω_i de l'ensemble fondamental Ω est associé un nombre positif p_i caractérisant ses chances d'apparition et appelé *probabilité*, tel que

$$p_1 + p_2 + \ldots + p_n + \ldots = \sum_{i\,:\,\omega_i \in \Omega} p_i = 1 \tag{4.2}$$

(de là, il résulte en particulier que $0 \leqslant p_i \leqslant 1$ quel que soit i).

Définition de la probabilité d'un événement. La probabilité de tout événement A se définit comme la somme des probabilités de tous les événements élémentaires composant A, c'est-à-dire si

l'on désigne par $P\{A\}$ la probabilité de A :

$$P\{A\} = \sum_{\omega_i \in A} P\{\omega_i\} = \sum_{i:\,\omega_i \in A} p_i. \tag{4.3}$$

De (4.3) et (4.2) on déduit en particulier que $0 \leqslant P\{A\} \leqslant 1$ de plus la probabilité de l'événement sûr est égale à 1, et celle de l'événement impossible à 0. Les autres notions et opérations sur les probabilités et les événements dérivent des quatre définitions exhibées (définitions d'une expérience aléatoire, d'une issue élémentaire, d'un événement aléatoire et de sa probabilité) et de l'axiome.

Donc, pour décrire exhaustivement le mécanisme de l'expérience aléatoire étudiée (dans le cas discret), il est nécessaire de se donner l'ensemble fondamental fini ou dénombrable Ω et d'associer à chaque issue élémentaire ω_i un nombre positif p_i ($\leqslant 1$) qui est la probabilité $P\{\omega_i\}$ d'apparition de ω_i, la correspondance $\omega_i \leftrightarrow p_i$ devant vérifier la relation de normalisation (4.2).

L'*espace probabilisé* est précisément la notion qui formalise cette description du mécanisme de l'expérience aléatoire. Se donner un espace probabilisé c'est se donner un ensemble fondamental Ω et y définir une correspondance du type

$$\omega_i \leftrightarrow p_i = P\{\omega_i\}. \tag{4.4}$$

Il est évident que la correspondance (4.4) peut être donnée de plusieurs façons : par des tables, des graphiques, des formules analytiques et, enfin, par des algorithmes.

Comment construire l'espace probabilisé correspondant aux conditions étudiées ? Si la concrétisation des notions d'expérience aléatoire, d'événement aléatoire, d'ensemble fondamental, et, dans le cas discret, de tout événement composé, ne pose en général pas de problèmes, la situation est différente pour ce qui est de la définition des probabilités $P\{\omega_i\}$ à partir des conditions du problème envisagé. A cet effet, on se sert de l'une des trois approches suivantes.

L'*approche a priori* du calcul des probabilités $P\{\omega_i\}$ consiste à analyser théoriquement les conditions spécifiques de l'expérience aléatoire étudiée. Dans de nombreux cas, cette analyse préliminaire permet de justifier théoriquement le procédé de définition des probabilités cherchées. Il est possible, par exemple, que l'ensemble fondamental soit constitué d'un nombre N fini d'éléments, les conditions de réalisation de l'expérience aléatoire considérée étant telles que les probabilités des N issues élémentaires sont égales (c'est notamment le cas dans le jet d'une pièce équilibrée, d'un dé parfaitement symétrique, dans le tirage d'une carte dans un jeu bien battu, etc.). D'après l'axiome (4.2), la probabilité de chaque événement élémentaire est alors égale à $1/N$. Ceci permet d'obtenir une recette simple pour le calcul de la probabilité de tout événement : si un événement A est composé de N_A événements élémentaires, la défini-

tion (4.3) nous donne

$$P\{A\} = \frac{N_A}{N}. \tag{4.3'}$$

La formule (4.3′) exprime que la probabilité d'un événement peut être définie *dans cette classe de situations* comme le rapport du nombre d'issues favorables (c'est-à-dire d'issues élémentaires composant cet événement) au nombre de toutes les issues possibles (ceci est la *définition classique de la probabilité*). Dans l'axiomatique moderne, la formule (4.3′) n'est pas la définition de la probabilité : elle n'est valable que dans le cas particulier où toutes les issues élémentaires sont équiprobables.

L'*approche fréquentiste a posteriori* du calcul des probabilités $P\{\omega_i\}$ est basée en fait sur la définition de la probabilité adoptée dans le cadre de la conception fréquentiste de la probabilité (cette conception est exposée avec plus de détails, par exemple, dans [40], [90]). D'après cette conception, la probabilité $P\{\omega_i\}$ *se définit* comme la limite de la fréquence relative d'apparition de l'issue ω_i lorsque le nombre n d'expériences aléatoires augmente indéfiniment, c'est-à-dire que

$$p_i = P\{\omega_i\} = \lim_{n\to\infty} \frac{m_n(\omega_i)}{n}, \tag{4.5}$$

où $m_n(\omega_i)$ est le nombre d'expériences aléatoires (parmi le nombre total n d'expériences aléatoires) dans lesquelles apparaît l'événement élémentaire ω_i. En conséquence, pour la définition pratique (approchée) des probabilités p_i on propose de prendre les fréquences relatives d'apparition de l'événement ω_i dans une longue série d'expériences aléatoires *). Un tel procédé *de calcul* des probabilités p_i ne contredit pas la conception (axiomatique) moderne de la théorie des probabilités, puisque cette dernière a été bâtie de telle sorte que l'analogue empirique de la probabilité objective réelle $P\{A\}$ de tout événement A est la fréquence relative de réalisation de cet événement dans une série de n épreuves indépendantes. Les probabilités sont définies de manière différente dans ces deux conceptions : dans l'approche fréquentiste la probabilité n'est pas une propriété objective *existant a priori de l'événement étudié, elle n'apparaît que si l'on procède à une expérience* ou à une observation ; d'où une con-

*) Le mathématicien allemand R. Mises, dont le nom est attaché au développement de la conception fréquentiste de la probabilité, estimait qu'à tout problème de probabilité correspondait nécessairement un processus réel (vérifiant les conditions introduites par lui de « collectif statistique ») et, de ce fait, qualifiait la théorie des probabilités de science des phénomènes du monde réel et non de science mathématique. La notion de « collectif » de Mises implique une apparition fréquente de l'événement (et de l'expérience respective) considéré, l'existence des limites (4.5) et l'indépendance de ces limites par rapport à la suite partielle d'expériences aléatoires choisies pour les calculer.

fusion entre les caractéristiques probabilistes théoriques (conditionnées par les conditions réelles d'« existence » du phénomène étudié) et leurs analogues empiriques. Comme le note Cramer « cette définition de la probabilité peut être comparée, par exemple, à la définition d'un point géométrique comme la limite d'une tache de craie de dimension indéfiniment décroissante, mais la géométrie axiomatique contemporaine n'a pas adopté une telle définition » ([40]). Nous ne nous attarderons pas sur les lacunes mathématiques de la conception fréquentiste de la probabilité. Nous indiquerons seulement les difficultés soulevées par le *calcul* des valeurs approchées p_i à l'aide des fréquences relatives $m_n(\omega_i)/n$. *Premièrement*, les conditions d'expérience aléatoire (c'est-à-dire les conditions d'ensemble statistique) dans laquelle on admet la tendance des fréquences relatives à se regrouper autour d'une valeur constante, ne peuvent être conservées indéfiniment et avec une précision élevée. Donc, pour estimer les probabilités p_i à l'aide des fréquences relatives $m_n(\omega_i)/n$, on n'a pas intérêt à prendre des séries trop longues (c'est-à-dire n assez grand), ce qui, soit dit en passant, fait perdre son sens à la limite (4.5). *Deuxièmement*, dans les situations donnant lieu à un nombre assez grand d'issues élémentaires (on sait que ce nombre peut être infini ou même, comme signalé au § 4.1, avoir la puissance du continu), même dans les séries d'épreuves arbitrairement longues nous aurons des issues *possibles* ω_i qui ne se produiront jamais ; quant aux valeurs approchées des probabilités acquises à l'aide des fréquences relatives des autres issues possibles, elles sont peu crédibles dans ces conditions.

Approche modélique a posteriori. Cette méthode de définition des probabilités $P\{\omega_i\}$, correspondant au système de conditions étudié, est actuellement de loin la plus répandue et la plus commode sur le plan pratique. Son principe est le suivant. D'une part, des lois de probabilité modèles (les lois binomiale, de Poisson, normale, exponentielle, etc., cf. § 6.1) ont été élaborées et analysées dans le cadre de l'approche *a priori*. De l'autre, l'analyste a entre les mains les résultats d'un nombre fini d'expériences aléatoires. Par ailleurs, par des méthodes spéciales de statistique mathématique (tests des paramètres inconnus, cf. chap. 8 et 9) il « ajuste » en quelque sorte les lois modèles à ses résultats (qui reflètent la réalité étudiée) et ne conserve que le ou les modèles qui ne contredisent pas ces résultats et leur correspondent le mieux dans une certaine optique.

Décrivons maintenant les principales opérations sur les probabilités des événements, qui découlent des définitions et axiomes adoptés plus haut.

Probabilité de la somme d'événements (théorème d'addition des probabilités). Formulons et prouvons la règle de calcul de la probabilité de la somme de deux événements A_1 et A_2. A cet effet, décomposons les ensembles des événements élémentaires constituant

A_1 et A_2 en deux parties:

$$A_1 = A_1^0 + A_{12};\ A_2 = A_2^0 + A_{12},$$

où A_k^0 est la différence symétrique de A_k et A_m (k, $m = 1, 2$; $k \neq m$), c'est-à-dire est constituée des événements élémentaires de A_k n'appartenant pas à A_m, et A_{12} est l'intersection de A_1 et A_2. La définition (4.3) et celle du produit des événements A_1 et A_2 nous donnent:

$$P\{A_1\} = \sum_{i:\,\omega_i \in A_1} p_i = \sum_{i:\,\omega_i \in A_1^0} p_i + \sum_{i:\,\omega_i \in A_{12}} p_i = P\{A_1^0\} + P\{A_1 \cdot A_2\}; \tag{4.6}$$

$$P\{A_2\} = \sum_{i:\,\omega_i \in A_2} p_i = \sum_{i:\,\omega_i \in A_2^0} p_i + \sum_{i:\,\omega_i \in A_{12}} p_i = P\{A_2^0\} + P\{A_1 \cdot A_2\}. \tag{4.7}$$

D'autre part, en vertu de la définition de la somme des événements $A = A_1 + A_2$ et de (4.3), on obtient

$$P\{A_1 + A_2\} = \sum_{i:\,\omega_i \in A_1^0} p_i + \sum_{i:\,\omega_i \in A_2^0} p_i + \sum_{i:\,\omega_i \in A_{12}} p_i. \tag{4.8}$$

Les relations (4.6), (4.7) et (4.8) entraînent la *formule d'addition des probabilités* (de deux événements):

$$P\{A_1 + A_2\} = P\{A_1\} + P\{A_2\} - P\{A_1 A_2\}. \tag{4.9}$$

La formule (4.9) peut être généralisée à un nombre quelconque de termes (cf., par exemple, [54], p. 105)

$$P\{A_1 + A_2 + \ldots + A_k\} = \sum_{i=1}^{k} P\{A_i\} - \Delta_1 + \Delta_2 - \ldots + (-1)^{k+1}\Delta_{k-1}, \tag{4.9'}$$

où Δ_m ($m = 1, 2, \ldots, k-1$) sont des sommes de probabilités de la forme

$$\Delta_m = \sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \ldots \sum_{i_{m+1}=1}^{k} P\{A_{i_1} \cdot A_{i_2} \ldots A_{i_{m+1}}\},$$

la sommation n'étant réalisée visiblement qu'à la condition que tous les $i_1, i_2, \ldots, i_{m+1}$ soient *distincts* et $i_1 < i_2 < \ldots < i_{m+1}$. Dans le cas particulier où le système $A_1, A_2, \ldots, A_k$ est composé d'événements *incompatibles*, tous les produits de la forme $A_{i_1} \cdot A_{i_2} \cdot \ldots \cdot A_{i_{m+1}}$ ($m \geqslant 1$) seront des événements vides (ou impossibles) et la formule (4.9') nous donne

$$P\{A_1 + A_2 + \ldots + A_k\} = P\{A_1\} + P\{A_2\} + \ldots + P\{A_k\}. \tag{4.9''}$$

Probabilité d'un produit d'événements (théorème de multiplication des probabilités). Probabilité conditionnelle. Etudions le cas où une condition posée *a priori* ou la fixation d'un événement déjà réalisé exclut une partie des événements élémentaires de l'espace probabilisé analysé. Ainsi, quand on analyse un ensemble de N articles composé de N_1 articles de première sorte, de N_2 de deuxième sorte, de N_3 de troisième sorte et de N_4 de quatrième sorte ($N_1 + N_2 + N_3 + N_4 = N$), on a affaire à un espace probabilisé composé des issues élémentaires ω_1, ω_2, ω_3 et ω_4 de probabilités respectives $p_1 = N_1/N$, $p_2 = N_2/N$, $p_3 = N_3/N$ et $p_4 = N_4/N$ (ω_i est l'événement qui consiste en ce que l'article prélevé au hasard est de la i-ième sorte). Supposons que les conditions de triage impliquent qu'à une certaine étape les articles de première sorte soient séparés de la population générale. Toutes les conclusions probabilistes (notamment, le calcul des probabilités des divers événements) nous aurons à les tirer à partir d'une population *tronquée* constituée uniquement des pièces de deuxième, troisième et quatrième sorte. Dans ces cas, on parle de *probabilités conditionnelles*, c'est-à-dire de probabilités calculées sous réserve qu'un certain événement se produise. Dans le cas considéré, cet événement est l'événement $B = \{\omega_2, \omega_3, \omega_4\}$, c'est-à-dire l'événement qui consiste en ce que toute pièce prélevée au hasard soit de deuxième, de troisième ou de quatrième sorte. Si donc l'on s'intéresse au calcul de la probabilité conditionnelle de l'événement A (sachant que l'événement B s'est réalisé) qui consiste, par exemple, en ce qu'une pièce prélevée au hasard est de deuxième ou de troisième sorte, il est évident que cette probabilité conditionnelle (désignée par $P\{A \mid B\}$) peut être définie comme suit :

$$P\{A \mid B\} = \frac{N_2+N_3}{N_2+N_3+N_4} = \frac{N_2+N_3}{N} : \frac{N_2+N_3+N_4}{N} = \frac{P\{AB\}}{P\{B\}}.$$

Il est aisé de voir sur cet exemple que le calcul des probabilités conditionnelles consiste en fait à passer à un autre espace fondamental tronqué par la condition B dans lequel le *rapport* des probabilités des événements élémentaires est le même que dans l'espace primitif, mais ces probabilités sont normées (au moyen d'une division par $P\{B\}$) de telle sorte que la condition (4.2) soit remplie dans l'espace tronqué. On aurait, certes, pu ne pas introduire la terminologie des probabilités conditionnelles et utiliser simplement le langage des probabilités habituelles (« inconditionnelles ») dans l'espace tronqué. Mais les notations en termes de probabilités de l'espace fondamental primitif sont utiles dans les cas où de par les conditions du problème nous devons constamment nous rappeler qu'il existe un espace primitif plus vaste.

Etablissons la formule de la probabilité conditionnelle dans le cas général. Soit B un événement (non vide) déjà réalisé et soit A un événement dont il faut calculer la probabilité conditionnelle

$P\{A \mid B\}$. L'espace fondamental tronqué $\widetilde{\Omega}$ n'est composé que des événements élémentaires figurant dans B et, par suite, leurs probabilités (compte tenu de la condition de normalisation (4.2)) sont définies par les formules *):

$$\widetilde{p}_i = P_{\widetilde{\Omega}}\{\omega_i\} = \begin{cases} \dfrac{P_\Omega\{\omega_i\}}{P_\Omega(B)} = \dfrac{p_i}{P_\Omega\{B\}} & \text{si } \omega_i \in B; \\ 0 & \text{si } \omega_i \notin B. \end{cases} \tag{4.10}$$

Par définition, la probabilité $P\{A \mid B\}$ est la probabilité de l'événement A dans l'espace probabilisé tronqué $\{\widetilde{\Omega}, \widetilde{p}\}$ et, donc, en vertu de (4.3) et (4.10):

$$P_\Omega\{A \mid B\} = P_{\widetilde{\Omega}}\{A\} = \sum_{\omega_i \in A} P_{\widetilde{\Omega}}\{\omega_i\} = \sum_{i\,:\,\omega_i \in AB} \left(\frac{p_i}{P_\Omega\{B\}}\right) = \frac{\sum\limits_{i\,:\,\omega_i \in AB} p_i}{P_\Omega\{B\}},$$

c'est-à-dire que

$$P_\Omega\{A \mid B\} = \frac{P_\Omega\{AB\}}{P_\Omega\{B\}} \tag{4.11}$$

ou, ce qui est équivalent,

$$P_\Omega\{AB\} = P_\Omega\{A \mid B\} \cdot P_\Omega\{B\}. \tag{4.11'}$$

Les formules (4.11) et (4.11') sont appelées respectivement *formule de la probabilité conditionnelle* et *règle de multiplication des probabilités*.

Rappelons encore que l'étude des probabilités conditionnelles d'événements différents se réalisant *sous une même condition* B revient à considérer des probabilités ordinaires dans un autre espace fondamental (tronqué) $\widetilde{\Omega} = B$ en recalculant les probabilités respectives des événements élémentaires à l'aide de la formule (4.10). Donc, *les théorèmes généraux et les opérations sur les probabilités restent tous valables pour les probabilités conditionnelles pourvu que celles-ci soient calculées pour la même condition.*

Indépendance des événements. On dit que deux événements A et B sont indépendants si

$$P\{AB\} = P\{A\} \cdot P\{B\}. \tag{4.12}$$

Cette définition est légitime. Pour nous en assurer revenons au théorème de multiplication des probabilités (4.11') et voyons dans quels cas la formule (4.12) en résulte. Il est évident que ceci a lieu lorsque la probabilité conditionnelle $P\{A \mid B\}$ est égale à la probabi-

*) L'indice affectant la probabilité indique dans quel espace est calculée cette probabilité.

lité inconditionnelle correspondante $P\{A\}$ ou lorsque la connaissance du fait que l'événement B s'est réalisé n'influe en aucune façon sur les chances d'apparition de l'événement A.

La définition de l'indépendance se généralise à un système de plus de deux événements de la manière suivante. On dit que des événements $A_1, A_2, \ldots, A_k$ sont mutuellement indépendants si pour tous deux, trois, quatre, etc., événements de cet ensemble sont valables les opérations de multiplication suivantes:

$$\begin{aligned} &P\{A_{i_1}\cdot A_{i_2}\} = P\{A_{i_1}\}\cdot P\{A_{i_2}\};\\ &P\{A_{j_1}\cdot A_{j_2}\cdot A_{j_3}\} = P\{A_{j_1}\}\cdot P\{A_{j_2}\}\cdot P\{A_{j_3}\};\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &P\{A_1\cdot A_2\ldots A_k\} = P\{A_1\}\cdot P\{A_2\}\ldots P\{A_k\}. \end{aligned} \tag{4.13}$$

Il est évident que dans la première ligne on a

$$C_k^2 = \frac{k(k-1)}{1\cdot 2}$$

(combinaisons de k éléments deux à deux) équations, dans la deuxième, C_k^3, etc. Au total, la formule (4.13) regroupe $C_k^2 + C_k^3 + \ldots + C_k^k = 2^k - k - 1$ conditions. Les C_k^2 conditions de la première ligne suffisent pour assurer l'*indépendance deux à deux* de ces événements. Et bien que indépendance deux à deux et indépendance dans l'ensemble ne recouvrent pas à strictement parler la même idée, leur différence présente un intérêt plutôt théorique que pratique: dans la pratique il n'existe probablement pas d'exemples importants d'événements indépendants deux à deux qui ne soient pas indépendants dans leur ensemble.

L'indépendance des événements facilite considérablement l'analyse des probabilités rattachées au système d'événements étudié. En effet, si dans le cas général il faut 2^k probabilités pour décrire les probabilités de toutes les combinaisons des événements du système $A_1, A_2, \ldots, A_k$, il en faut seulement k, en l'occurrence les probabilités $P\{A_1\}, P\{A_2\}, \ldots, P\{A_k\}$, si ces événements sont deux à deux indépendants.

Les événements indépendants se rencontrent fréquemment en pratique: ils se produisent dans les expériences (observations) qui sont effectuées indépendamment l'une de l'autre au sens physique habituel.

C'est précisément l'indépendance des issues dans une série de quatre jets consécutifs de dé qui a permis (grâce à (4.13)) de calculer la probabilité de non-apparition (dans aucun de ces jets) du six dans le numéro 2.2.1. En effet, si A_i est l'événement consistant en la non-apparition du six dans le i-ième jet ($i = 1, 2, 3, 4$), on obtient,

en tenant compte du fait que $P\{A_i\} = 5/6$ pour tous les $i = 1, 2, 3, 4$.

$$P\{A_1 \cdot A_2 \cdot A_3 \cdot A_4\} = \\ = P\{A_1\} \cdot P\{A_2\} \cdot P\{A_3\} \cdot P\{A_4\} = (5/6)^4 = 625/1296.$$

Formule des probabilités totales. Souvent le calcul direct de la probabilité d'un événement A est difficile, voire impossible, alors que le calcul (ou la donnée) des probabilités conditionnelles du même événement A (sous des conditions différentes) ne pose pas de problème. Si les conditions $B_1, B_2, \ldots, B_k$, sous lesquelles sont connues (ou facilement calculables) les probabilités conditionnelles d'un événement A, forment un système complet d'événements (cf. n° 4.1.2) on peut calculer la probabilité $P\{A\}$ à l'aide de la relation

$$P\{A\} = P\{A \mid B_1\} P\{B_1\} + P\{A \mid B_2\} \cdot P\{B_2\} + \ldots \\ \ldots + P\{A \mid B_k\} \cdot P\{B_k\}, \quad (4.14)$$

dite *formule des probabilités totales.*

Pour prouver la formule (4.14), on remarquera que les événements élémentaires composant l'événement A peuvent être divisés k groupes disjoints qui sont chacun l'intersection de A avec un événement B_i (ceci résulte directement du fait que la somme des événements $B_1, B_2, \ldots, B_k$, qui sont deux à deux disjoints, recouvre l'espace fondamental tout entier), c'est-à-dire que

$$A = AB_1 + AB_2 + \ldots + AB_k.$$

En se servant ensuite du théorème d'addition des probabilités (4.9″) (appliqué aux événements incompatibles que sont les événements $AB_1, AB_2, \ldots, AB_k$) et en calculant la probabilité de chaque produit AB_i à l'aide de la formule (4.11′), on retrouve la formule (4.14).

Formule de Bayes. Commençons par le problème suivant. Dans un dépôt sont stockés des instruments construits par trois usines: 20 % de ces appareils proviennent de l'usine n° 1, 50 % de l'usine n° 2 et 30 % de l'usine n° 3. Les probabilités qu'un appareil tombe en panne durant le délai de garantie sont de 0,2, 0,1 et 0,3 si cet appareil sort respectivement de l'usine n° 1, n° 2 et n° 3. Un appareil sans label est tombé en panne (pendant le délai de garantie). Quelle est l'usine qui a le plus de chance de l'avoir produit? Quelle est cette probabilité? Si l'on désigne par A_i l'événement « un appareil prélevé au hasard provient de l'usine i » ($i = 1, 2, 3$), B l'événement « un appareil pris au hasard tombe en panne », le problème posé ci-dessus se ramène de toute évidence au calcul des probabilités conditionnelles suivantes: $P\{A_i \mid B\}$ sachant que $P\{A_1\} = 0{,}2$; $P\{A_2\} = 0{,}5$; $P(A_3\} = 0{,}3$; $P\{B \mid A_1\} = 0{,}2$; $P\{B \mid A_2\} = 0{,}1$; $P\{B \mid A_3\} = 0{,}3$. Les événements A_1, A_2 et A_3 formant un système complet, la formule de la probabilité conditionnelle (4.11)

nous donne

$$P\{A_i \mid B\} = \frac{P\{A_iB\}}{P\{B\}}. \tag{4.15}$$

D'après le théorème de multiplication des probabilités (4.11'), le numérateur de (4.15) peut être mis sous la forme

$$P\{A_iB\} = P\{BA_i\} = P\{B \mid A_i\}\cdot P\{A_i\}, \tag{4.16}$$

et, d'après la formule des probabilités totales (4.14), le dénominateur devient

$$P\{B\} = P\{B \mid A_1\}\cdot P\{A_1\} + P\{B \mid A_2\}\cdot P\{A_2\} + \\ + P\{B \mid A_3\}\cdot P\{A_3\}. \tag{4.17}$$

En portant (4.16) et (4.17) dans (4.15), on trouve

$$P\{A_i \mid B\} = \frac{P\{A_i\}\cdot P\{B \mid A_i\}}{\sum\limits_{j=1}^{3} P\{B \mid A_j\}\cdot P\{A_j\}}. \tag{4.18}$$

Cette formule nous conduit immédiatement aux probabilités cherchées :

$$P\{A_1 \mid B\} = \frac{0{,}2\cdot 0{,}2}{0{,}2\cdot 0{,}2+0{,}1\cdot 0{,}5+0{,}3\cdot 0{,}3} = \frac{0{,}04}{0{,}18} = 0{,}222\,;$$

$$P\{A_2 \mid B\} = \frac{0{,}5\cdot 0{,}1}{0{,}18} = \frac{0{,}05}{0{,}18} = 0{,}278\,;$$

$$P\{A_3 \mid B\} = \frac{0{,}3\cdot 0{,}3}{0{,}18} = 0{,}500.$$

Donc, l'appareil défectueux a le plus de chance de sortir de l'usine n° 3.

La démonstration de la formule (4.18) dans le cas d'un système complet de k événements reprend *in extenso* celle de la formule (4.18). Sa forme générale

$$P\{A_i \mid B\} = \frac{P\{A_i\}\cdot P\{B \mid A_i\}}{\sum\limits_{j=1}^{k} P\{B \mid A_j\}\cdot P\{A_j\}} \tag{4.19}$$

s'appelle *formule de Bayes.*

4.2. Espace probabilisé continu (axiomatique de Kolmogorov)

4.2.1. Spécificité du cas général (continu) de l'espace probabilisé. Nous avons évoqué précédemment des cas d'ensembles fondamentaux Ω *plus que dénombrables.* Ainsi, l'espace fondamental a la puissance du *continu* si à chaque issue élémentaire ω_i est associé un ou plusieurs caractères de l'objet étudié, mesurés par des unités *continues* (de temps, longueur, poids, température, pression, etc.). On pourrait certes objecter que la *précision* des mesures étant *limitée,*

l'ensemble fondamental réel sera au plus dénombrable. Or, d'une part, cette précision ne cesse de s'améliorer, donc la structure de l'espace probabilisé discret se modifiera en conséquence. De l'autre, l'étude de modèles continus reflétant la nature physique de l'événement analysé *accroît les possibilités analytiques de la théorie*, fournit à l'analyste un outil mathématique plus puissant : il suffit de comparer les possibilités de la sommation et de l'intégration, les possibilités des équations aux différences et des équations différentielles, etc.

Comment s'opère le passage du cas discret au cas continu lors de la construction d'une théorie des possibilités rigoureusement mathématique? L'extension automatique du schéma de construction de l'espace probabilisé discret (cf. § 4.1) au cas continu est impossible. Une différence fondamentale entre le cas continu et le cas discret est que *d'une façon générale on ne peut*, comme ceci a été fait dans un espace probabilisé discret, *déclarer toute partie de l'ensemble fondamental* Ω *événement aléatoire*, c'est-à-dire un événement observable dans le cadre de l'expérience aléatoire envisagée. En d'autres termes, dans un espace probabilisé continu les parties de l'ensemble fondamental Ω qui jouissent de la propriété d'être observables sont dites *mesurables*, les autres, *non mesurables*.

Citons un *exemple de non-mesurabilité* de parties de l'ensemble fondamental due aux capacités réduites des instruments d'observation utilisés. Supposons qu'un astronome dispose d'un télescope lui permettant de fixer la position des seules étoiles dont la luminosité excède un certain seuil. Pour ensemble fondamental nous prendrons l'ensemble des positions de *toutes* les étoiles (et non pas uniquement celles qui sont accessibles à l'astronome) dans l'espace. Il est évident que l'ensemble Ω est plus que dénombrable (a la puissance du continu). Pour l'astronome, sous réserve qu'il n'utilise pas les acquis de l'astronomie, sont non mesurables (*non vérifiées expérimentalement*) toutes les assertions concernant les étoiles dont la luminosité est inférieure au seuil. Ceci étant, nous savons qu'avec des instruments plus puissants, la prise en compte du mouvement de la Terre autour du Soleil ou la mise en œuvre d'autres méthodes de l'astronomie moderne, une partie de ces assertions pourrait être vérifiée. La notion de mesurabilité permet dans le cas présent de faire une nette distinction entre les assertions vérifiables et celles non encore en l'état de l'être sur la structure de l'espace probabilisé. Quant à l'ensemble fondamental Ω, il reste invariant en tant qu'état de la nature.

Ainsi, il est naturel d'appeler événements aléatoires les seuls sous-ensembles A de l'ensemble fondamental Ω à propos desquels il est possible de dire s'ils ont eu lieu ou non dans le cadre d'une expérience, puisque dans ce cas seulement on peut parler de la fréquence relative de leur apparition dans une série de n expériences, donc, de la *probabilité* $P\{A\}$.

Cette particularité du cas continu implique vraisemblablement l'introduction de définitions et axiomes supplémentaires sur les événements aléatoires, les opérations sur eux et sur leurs probabilités. Cette procédure est notamment appliquée lors de la construction axiomatique (ensembliste) de la théorie des probabilités dont le premier exposé complet et rigoureux est dû à Kolmogorov ([81]) *).

4.2.2. Evénements aléatoires, probabilités et opérations (approche axiomatique de Kolmogorov). Définissons la partie des sous-ensembles de l'ensemble fondamental Ω qui contient des *événements*. La procédure de définition d'un événement aléatoire A dans le cas continu est identique à celle adoptée pour le cas discret. Mais, si dans le cas discret, il nous a suffit de définir comme notions premières les événements élémentaires $\omega_1, \omega_2, \ldots, \omega_N, \ldots$ (et tout sous-ensemble de l'ensemble fondamental était un événement), dans le cas continu, dans chaque situation concrète nous devons (à partir de considérations physiques) définir en plus de Ω une catégorie de sous-ensembles de Ω qui sont visiblement des événements. Ensuite, tout événement aléatoire A se définit comme une combinaison des sous-ensembles de cette catégorie.

Définition d'un événement aléatoire. Considérons un système (fini ou dénombrable) de sous-ensembles $A_1, A_2, \ldots, A_n, \ldots$ de l'ensemble fondamental Ω, chacun de ces sous-ensembles étant un événement. Alors, l'ensemble Ω, les complémentaires $\overline{A}_i = \Omega - A_i$, $\varnothing = \Omega - \Omega$ et la somme $A = A_1 + A_2 + \ldots + A_n + \ldots$ sont aussi des événements (on en déduit immédiatement que le produit $\Pi = A_1 \cdot A_2 \cdot \ldots \cdot A_n \cdot \ldots$ est aussi un événement, puisque son complémentaire $\overline{\Pi} = \overline{A}_1 + \overline{A}_2 + \ldots$ l'est d'après la définition donnée).

Dans la suite, on désignera par **C** le système des sous-ensembles de l'ensemble fondamental Ω qui sont des événements.

Axiome. A chaque événement A du système **C** est rattaché un nombre $p(A) = P\{A\}$ compris entre 0 et 1 appelé probabilité de l'événement A, tel que

a) $p(\Omega) = P\{\Omega\} = 1$;

b) si des événements $A_1, A_2, \ldots, A_n, \ldots$ sont incompatibles, alors

$$P\{A_1 + A_2 + \ldots + A_n + \ldots\} = \\ = P\{A_1\} + P\{A_2\} + \ldots + P\{A_n\} + \ldots$$

*) La première interprétation d'un événement aléatoire, comme un ensemble, et de sa probabilité, comme une mesure de cet ensemble, est probablement l'œuvre du mathématicien polonais A. Lomnicki (Lomnicki A. *Nouveaux fondements du calcul des probabilités* — Fundam. Math., 1923, 4). Cependant, la « théorie des probabilités comme théorie mathématique exacte a été construite dans un volume convenable pour la première fois par A. Kolmogorov » ([138], p. 11).

De cet axiome, on déduit en particulier la relation suivante entre les probabilités des événements A et $\overline{A}$:

$$P\{\overline{A}\} = 1 - P\{A\}.$$

La propriété axiomatique b) des probabilités a été formulée et *prouvée* dans le cas discret sous la forme du théorème d'addition des probabilités.

De façon analogue, le théorème de multiplication des probabilités (qui a été déduit des définitions et axiomes du cas discret) *est adopté par définition* dans le cas continu.

Définition de la probabilité conditionnelle. La probabilité conditionnelle $P\{A \mid B\}$ d'un événement A sachant qu'un événement B s'est réalisé est donnée par la formule

$$P\{AB\} = P\{A \mid B\} \cdot P\{B\}.$$

Les **opérations sur les événements et leurs probabilités** et, en particulier, la formule des probabilités totales (4.14), la formule de Bayes (4.18), la définition de l'indépendance pour le système d'événements (4.12), (4.13) et les autres formules prouvées dans le cas discret sont valables (et peuvent être établies) dans le cas d'un espace probabilisé continu.

Récapitulons. Pour décrire complètement le mécanisme d'une expérience aléatoire étudiée dans le cas continu, c'est-à-dire pour définir l'espace probabilisé, il faut : 1) décrire l'ensemble des événements élémentaires Ω ; 2) décrire le système $\mathbf{C}$ des sous-ensembles mesurables de Ω ou des sous-ensembles *observables* (c'est-à-dire des événements) ; 3) à chaque événement A du système $\mathbf{C}$ associer un nombre positif $P\{A\}$ appelé probabilité de l'événement A, cette correspondance vérifiant les conditions a) et b) et l'axiome (il est évident que l'application P est une fonction d'ensemble numérique définie sur les sous-ensembles du système $\mathbf{C}$; de telles applications sont dites *mesures de probabilité* définies sur le système de sous-ensembles $\mathbf{C}$). Donc, si dans le cas discret deux symboles : Ω et p, ont suffi à décrire l'espace probabilisé, dans le cas continu, il en faut trois : Ω, $\mathbf{C}$, P.

L'exemple des observations astronomiques conforte en quelque sorte le caractère naturel des notions de sous-ensemble mesurable (« observable») et non mesurable (« non observable ») de l'ensemble des événements élémentaires Ω. Mais tout modèle, toute théorie, y compris la conception axiomatique de la théorie des probabilités, n'est qu'une forme de représentation *approchée* de la réalité, une forme qui n'est pas exempte de défauts. L'exemple suivant est destiné à prévenir le lecteur contre toute surestimation des vertus de l'approche axiomatique.

Les *sondages d'experts* sont utilisés sous des formes différentes pour les prévisions socio-économiques et scientifico-techniques à long terme. Une forme consiste à demander à chaque expert d'évaluer subjectivement la probabilité de réalisation d'un événement *). Si l'on aborde la modélisation de ce processus du point de vue de l'école subjectiviste et que, par conséquent, l'on interprète chaque expert comme un « instrument de mesure », on peut définir la notion d'expérience aléatoire comme le résultat de la réponse de l'expert à la question posée. Dans ce cas, l'espace des issues élémentaires Ω doit évidemment être composé de tous les points de l'intervalle $[0, 1]$. En construisant le système C des « observables » de l'espace Ω, il eût été naturel d'exiger *a priori* que tout intervalle $\Delta = [c_1, c_2]$, contenu dans l'intervalle $[0, 1]$ (c'est-à-dire que $0 \leqslant c_1 < c_2 \leqslant 1$), appartienne à la catégorie des *événements* (autrement dit, pour tout intervalle $\Delta = [c_1, c_2]$ doit être définie la probabilité $P\{\Delta\} = P\{[c_1, c_2]\}$ que la réponse chiffrée d'un expert pris au hasard appartienne à cet intervalle). Mais alors, d'après la définition d'un événement aléatoire dans le cas continu, seront des événements non seulement les intervalles, mais la somme et le produit d'un nombre dénombrable, d'entre eux, ainsi que leurs complémentaires (c'est-à-dire les événements contraires). Donc, si l'on prend un point quelconque $c \in [0, 1]$ et que l'on considère la suite d'intervalles $\Delta_n = [c - 1/n, c + 1/n]$, on constate que ce *point* doit être un événement, car il est le produit d'un nombre dénombrable d'intervalles Δ_n. Ainsi tout point de $[0, 1]$ est un événement. L'ensemble des points rationnels est dénombrable. Donc, c'est un événement aussi. Il en est de même de l'ensemble des points irrationnels, car complémentaire de celui des rationnels. Mais il est peu probable que du *point de vue physique* l'appartenance à l'ensemble des rationnels et des irrationnels puisse être considérée comme un observable (donc comme un événement physiquement discernable). Cet exemple montre que l'*approche axiomatique de la théorie des probabilités peut conduire à des conclusions dont l'interprétation physique est erronée.* La construction de l'espace probabilisé n'a pas été menée à son terme ici, puisque le procédé de calcul des probabilités sur les intervalles, soit $P\{\Delta\} = P\{[c_1, c_2]\}$, n'a pas été défini. La définition axiomatique de ces probabilités dépend aussi de la nature des conditions caractérisant notre expérience aléatoire. Ainsi, si l'on s'imagine être dans la *situation la moins favorable* (l'événement qui nous préoccupe est tellement éloigné et incertain ou les experts sondés tellement incompétents que leurs réponses sont *grosso modo* uniformément distribuées sur l'intervalle $[0, 1]$), alors il est naturel d'admettre que les probabilités $P\{\Delta\}$ dépendront seulement de la lon-

*) Cette forme d'expertise est utilisée, par exemple, pour composer des scénarios pour le développement socio-économique et industriel d'un pays.

gueur de l'intervalle Δ et pas de sa position, et de les définir par conséquent à l'aide des relations *).

$$P\{[c_1, c_2]\} = c_2 - c_1. \tag{4.20}$$

Il est immédiat de vérifier que ces probabilités satisfont les propriétés a) et b) de l'axiome.

Signalons dans le cadre de cet exemple une distinction fondamentale entre un grand nombre d'espaces probabilisés continus et d'espaces probabilisés discrets: la probabilité de réalisation de tout événement élémentaire ω est nulle ici; or la probabilité $P\{\Delta\}$, où Δ est un intervalle aussi petit que l'on veut, sera toujours strictement positive (ceci résulte directement de (4.20)). Donc, cet exemple nous met pour la première fois en face de la situation, semble-t-il paradoxale, d'*événements ω qui bien que possibles n'en possèdent pas moins une probabilité nulle.* De façon respective, les *événements $\bar{\omega} = \Omega - \omega$ complémentaires d'événements ω de probabilité nulle ne peuvent être appelés certains bien que leur probabilité soit égale à l'unité.* A propos des événements $\bar{\omega}$, on dit qu'ils ont lieu « presque sûrement ». Une étude plus poussée montre que dans un espace probabilisé continu, ces situations ne sont pas paradoxales. Pour comprendre cette idée, on pourrait songer à un solide qui a une certaine masse, alors que les points qui le composent sont impondérables. Il est évident que dans cette analogie le solide joue le rôle d'un événement, le point, le rôle d'une issue élémentaire, la masse, le rôle de la probabilité.

Conclusions

1. Les fondements de l'appareil mathématique de la théorie des probabilités sont: les *notions* d'expérience aléatoire, de son issue possible et d'ensemble des événements élémentaires; l'*axiome* d'existence et de normalisation des probabilités des événements élémentaires; la *définition* d'un événement aléatoire et de la méthode de calcul de sa probabilité.

2. La théorie des probabilités moderne s'édifie *axiomatiquement.* Ceci étant, pour construire l'espace probabilisé discret, c'est-à-dire pour décrire mathématiquement le mécanisme d'une expérience aléatoire dont l'ensemble des issues élémentaires est fini ou dénombrable, il suffit de postuler un axiome (sur l'existence et la normalisation des probabilités des issues élémentaires) et une définition (sur la méthode de calcul de la probabilité de tout événement).

3. Les termes « mécanisme d'une expérience aléatoire », « système de conditions induisant l'ensemble statistique étudié » et « espace probabilisé » sont synonymes et peuvent être définis mathématiquement par la description de toutes les issues élémentaires

*) Sur les trois approches possibles (*a priori*, fréquentiste *a posteriori* et modélique *a posteriori*) de l'élaboration et de la justification de telles hypothèses relativement à la nature des probabilités définies axiomatiquement voir n° 4.1.3.

possibles et par la donnée de la loi qui associe à chacune d'elles la probabilité de son apparition (sous une forme analytique, tabulaire, graphique ou algorithmique).

4. La *principale difficulté* soulevée par la construction de l'espace probabilisé correspondant au système de conditions étudié réside dans la définition des probabilités des événements élémentaires. Des trois approches de résolution de ce problème — l'approche *a priori*, l'approche *fréquentiste a posteriori* et l'approche *modélique a posteriori*, — la dernière est la plus efficace et la plus facile à réaliser en pratique.

5. Les principales opérations dans l'espace probabilisé discret sont définies par les *théorèmes d'addition et de multiplication des probabilités, les formules des probabilités totales et de Bayes.*

6. Contrairement au cas discret, dans le cas continu l'ensemble des événements élémentaires Ω peut contenir des sous-ensembles non observables (non mesurables). Ces sous-ensembles ne peuvent être qualifiés d'événements, car, si A était un événement, il faudrait être en mesure de dire s'il est le résultat ou non d'une expérience (en ce sens il est « observable »); alors seulement on pourrait parler de la fréquence relative de sa réalisation dans une série d'expériences et, partant, de sa probabilité $P\{A\}$.

7. Cette particularité de l'espace probabilisé continu implique l'introduction de définitions et d'axiomes subsidiaires régissant les événements aléatoires et les opérations sur leurs probabilités. La théorie axiomatique des probabilités (dont l'exposé complet et rigoureux a été donné pour la première fois en 1933 par Kolmogorov) construit l'espace probabilisé continu en partant de la définition d'un événement aléatoire (moyennant une énumération des opérations ensemblistes possibles sur les sous-ensembles qui sont *a priori* des événements) et de l'axiome des probabilités, probabilités qui sont définies comme des fonctions d'ensembles-événements prenant leurs valeurs dans l'intervalle $[0, 1]$. Cette conception ne contredit pas la méthode de construction de l'espace probabilisé discret proposé plus haut (elle comprend cette méthode comme cas particulier et préserve, par conséquent, toutes les opérations sur les probabilités et les événements) et permet d'interpréter du point de vue physique la probabilité d'un événement comme la fréquence relative de son apparition dans une assez longue série d'expériences.

8. La conception axiomatique de la théorie des probabilités peut dans certains cas, comme d'ailleurs tout autre modèle, conduire à des conclusions dont l'interprétation physique est mauvaise.

9. L'espace probabilisé continu peut à l'inverse de l'espace discret contenir des événements de probabilité nulle. Les complémentaires de ces événements, bien qu'ils ne puissent être qualifiés de certains, n'en possèdent pas moins une probabilité égale à 1 (ces événements sont presque sûrs).

CHAPITRE 5

VARIABLES ALÉATOIRES

5.1. Définition et exemples de variables aléatoires

Les exemples d'expériences aléatoires (exemples 4.1 à 4.7 du n° 4.1.2) montrent que dans la plupart des cas le résultat d'une expérience aléatoire se laisse décrire par un ou plusieurs nombres. Dans les exemples 4.2, 4.5 et 4.7, ces résultats désignent respectivement le nombre de points de la face supérieure d'un dé ; le nombre de pièces défectueuses dans un échantillon de taille N ; le nombre de pannes d'une chaîne desservie par une équipe prise au hasard. Dans l'exemple 4.3 (série de 4 jets de dé), le résultat de chaque expérience aléatoire se représente par quatre chiffres ou un vecteur à quatre dimensions, et dans l'exemple 4.6 (test de deux échantillons de taille respective N_1 et N_2), par un couple de nombres ou un *vecteur à deux dimensions.* Même dans les exemples 4.1 et 4.4 qui, de prime abord, n'ont aucun lien avec des nombres, il est possible de coder les issues en attribuant par exemple le nombre 0 à « pile » (exemple 4.1) et à « bon » (exemple 4.4), et le nombre 1 respectivement à « face » et « mauvais ». En poursuivant nos exemples et en considérant, dans le cadre ensembliste, l'analyse d'un ou de plusieurs caractères (exprimés par des nombres ou codés) des objets étudiés, on arrive à un schéma général dans lequel la notion d'expérience aléatoire se traduit par l'enregistrement d'un ensemble de nombres

$$\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})', \quad p \geqslant 1. \tag{5.1}$$

Les valeurs de ξ obtenues au terme de l'expérience envisagée dépendent d'une foule de facteurs aléatoires et sont définies de façon unique par l'issue élémentaire ω de cette expérience (or cela signifie que ξ est une fonction numérique (pour $p = 1$) ou vectorielle (pour $p \geqslant 2$) de l'argument ω). Nous sommes ainsi conduits à la définition suivante de la variable aléatoire.

On appelle *variable aléatoire une quantité scalaire* ($p = 1$) *ou vectorielle* ($p \geqslant 2$) *mesurable recouvrant un certain sens physique, dont les valeurs* (*les composantes*) *sont affectées d'une dispersion incontrôlable lors de la répétition de l'expérience* (*de l'observation, du processus*). *On pourrait dire aussi qu'une variable aléatoire* ξ *est une fonction définie sur un ensemble d'événements élémentaires, c'est-à-dire que* $\xi = \xi(\omega)$.

Exemple 5.1. Le tableau 5.1 représente encore un exemple de variable aléatoire *vectorielle* (ou *multidimensionnelle*) avec la forme générale correspondante d'enregistrement d'une série d'observations.

Les symboles de la variable aléatoire (variable qui en fait représente la *liste des caractères* analysés et se note le plus souvent par une lettre grecque : ξ, η, ζ, ν, etc.) diffèrent des notations adoptées pour

Tableau 5.1

	Caractères analysés des objets ξ	Objets analysés (familles)					
		1	2	...	j	...	n
Caractères de l'« état » de l'objet (caractères socio-démographiques et économiques de la famille)	$\xi^{(1)}$ milieu social	$x_1^{(1)}$	$x_2^{(1)}$		$x_j^{(1)}$		$x_n^{(1)}$
	$\xi^{(2)}$ profession du chef de famille	$x_1^{(2)}$	$x_2^{(2)}$		$x_j^{(2)}$		$x_n^{(2)}$
	$\xi^{(3)}$ conditions de logement	$x_1^{(3)}$	$x_2^{(3)}$		$x_j^{(3)}$		$x_n^{(3)}$
	$\xi^{(4)}$ nombre de membres de la famille	$x_1^{(4)}$	$x_2^{(4)}$		$x_j^{(4)}$		$x_n^{(4)}$
	$\xi^{(5)}$ nombre d'enfants	$x_1^{(5)}$	$x_2^{(5)}$		$x_j^{(5)}$		$x_n^{(5)}$
	$\xi^{(6)}$ revenu moyen par membre	$x_1^{(6)}$	$x_2^{(6)}$		$x_j^{(6)}$		$x_n^{(6)}$
Caractères du « comportement » de l'objet (structure de la consommation de la famille, roubles, par membre, par mois)	$\xi^{(7)}$ dépenses pour la nourriture	$x_1^{(7)}$	$x_2^{(7)}$		$x_j^{(7)}$		$x_n^{(7)}$
	$\xi^{(8)}$ dépenses pour les articles d'usage courant	$x_1^{(8)}$	$x_2^{(8)}$		$x_j^{(8)}$		$x_n^{(8)}$
	$\xi^{(9)}$ dépenses pour les articles de luxe et à long usage	$x_1^{(9)}$	$x_2^{(9)}$		$x_j^{(9)}$		$x_n^{(9)}$
	$\xi^{(10)}$ dépenses pour les services	$x_1^{(10)}$	$x_2^{(10)}$		$x_j^{(10)}$		$x_n^{(10)}$
	$\xi^{(11)}$ autres dépenses, épargne incluse	$x_1^{(11)}$	$x_2^{(11)}$		$x_j^{(11)}$		$x_n^{(11)}$

les valeurs *observées*. Dans le tableau 5.1 et dans la suite les valeurs observées sont désignées à des fins d'unification par des lettres latines minuscules (la lettre x dans le tableau 5.1) affectées d'un indice supérieur pour le caractère analysé et d'un indice inférieur

pour le numéro de l'expérience ou de l'objet étudié; mais dans tous les cas il faut avoir présent à l'esprit que ces symboles « recouvrent » des *valeurs numériques* des caractères correspondants ou leurs *valeurs codées*. Ainsi, dans le tableau 5.1, de toute évidence les trois premières lignes seront composées de valeurs codées, les huit suivantes, de valeurs numériques revêtant une signification physique claire mesurées dans une certaine échelle.

5.2. Valeurs possibles et observées d'une variable aléatoire

Etant donné qu'une variable aléatoire ξ est une fonction définie sur l'ensemble des issues élémentaires, ses *valeurs possibles* et leur nombre dépendent de la structure de l'espace probabilisé correspondant Ω: à chaque événement élémentaire ω est associée une valeur éventuelle de ξ. Donc, il y a autant d'événements élémentaires que de valeurs de $\xi(\omega)$. Ainsi, dans les exemples du n° 4.1.2 nous avons: pour les exemples 4.1 et 4.4, en tout deux valeurs correspondant aux issues élémentaires « pile » et « face » (pour le premier exemple) et « bon » et « mauvais » (pour le second); pour l'exemple 4.2, les six premiers nombres entiers; pour l'exemple 4.3, tous les quadruplets formés avec les nombres 1 à 6 (on a vu que ces quadruplets étaient au nombre de $6^4 = 1296$); pour l'exemple 4.5, tous les entiers positifs compris entre 0 et N; pour l'exemple 4.6, tous les couples de nombres dont le premier prend les valeurs entières comprises entre 0 et N_1, et le second, entre 0 et N_2 (il est évident que le nombre total de ces couples est égal à $(N_1 + 1)(N_2 + 1)$); pour l'exemple 4.7, tous les entiers positifs. Il est certes impossible de « dénombrer » toutes les valeurs prises par une variable aléatoire définie sur l'espace *continu* des événements élémentaires Ω: ces valeurs forment un ensemble qui a la puissance du continu. C'est précisément ce type de variables aléatoires qui est représenté par les composantes $\xi^{(6)}$ à $\xi^{(11)}$ de ξ dans l'exemple 5.1.

Il y a lieu de faire une distinction entre les *valeurs théoriques* de la variable aléatoire (ces valeurs sont désignées par $x_1^0, x_2^0, \ldots, x_n^0, \ldots$ dans le cas discret et simplement par x, dans le cas continu et les *valeurs empiriques*, c'est-à-dire *observées dans les expériences* (ces valeurs seront notées $x_1, x_2, \ldots, x_n$) *).

5.3. Types de variables aléatoires

La classification générale des divers types de variables aléatoires est donnée par le schéma de la figure 5.1.

Si le résultat de l'expérience (de l'observation) est mesuré par

*) Si la variable aléatoire étudiée est vectorielle (cf. § 5.1), ces notations restent en vigueur, mais les lettres latines minuscules x_i^0, x, x_i seront remplacées par les majuscules correspondantes X_i^0, X, X_i.

un seul nombre (exemples 4.1, 4.2, 4.4, 4.5 et 4.7 du n° 4.1.2; cf. aussi le cas $p = 1$ dans (5.1)), la variable aléatoire correspondante est dite *scalaire* ou *à une dimension*. Si ce résultat est décrit par un *ensemble de nombres* (exemples 4.3, 4.6 et 5.1, ainsi que le cas $p \geqslant 2$ dans (5.1)), la variable aléatoire correspondante s'appelle *vectorielle* ou *multidimensionnelle*.

Une variable aléatoire scalaire est *discrète* ou *continue* selon qu'elle est définie sur un espace des événements élémentaires discret

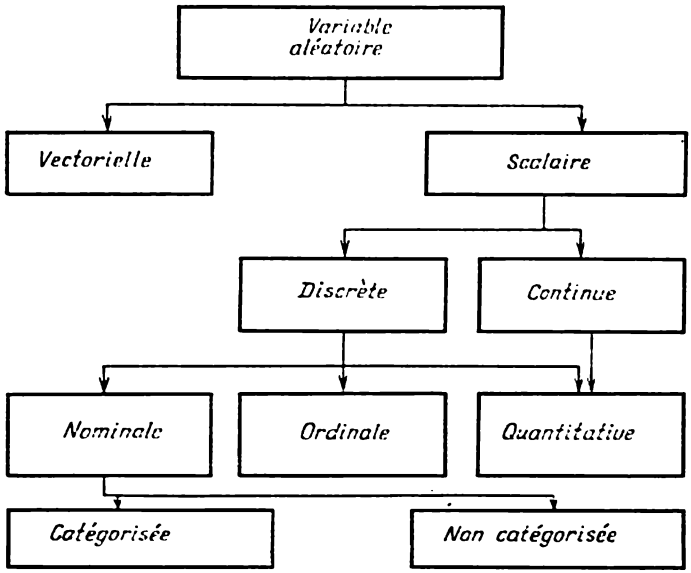


Fig. 5.1. Schéma général de classification des principaux types de variables aléatoires

ou continu. Il est évident que dans tous les exemples 4.1 à 4.7 de même que dans les cinq premières composantes de l'exemple 5.1 (tableau 5.1), nous avons affaire à des variables aléatoires discrètes. Nous avons déjà signalé que certains chercheurs arguent de la *précision limitée des mesures* pour préconiser d'utiliser uniquement des espaces probabilisés discrets, et respectivement uniquement des variables aléatoires discrètes. En effet, même quand on mesure des quantités continues (longueur, poids, température, pression, etc.), il existe toujours une unité discernable, une sorte de quantum indivisible, dépendant du pouvoir de résolution des instruments utilisés, qui sert à représenter l'objet étudié. Mais les possibilités analytiques des modèles mathématiques continus, les résultats de leur utilisation indiquent qu'ils sont un appareil efficace non seulement pour les variables aléatoires continues, mais aussi pour les variables discrè-

tes dont les ensembles des valeurs sont suffisamment riches (cf. exemple du § 6.1 et fig. 5.6).

Les variables aléatoires discrètes scalaires se divisent en quantitatives, ordinales et nominales selon leur nature et leur destination.

La *variable aléatoire quantitative* permet de mesurer dans une certaine échelle le degré de présence du caractère analysé dans l'objet sondé (voir les exemples 4.2, 4.3, 4.5, 4.6, 4.7, ainsi que les composantes $\xi^{(4)}$ à $\xi^{(11)}$ dans l'exemple 5.1).

La *variable aléatoire ordinale* permet d'ordonner les objets d'après le degré de présence du caractère analysé. L'analyste fait appel aux variables aléatoires ordinales dans les cas où l'échelle dans laquelle il doit mesurer le degré de présence du caractère analysé n'existe pas ou lui est inconnue. Dans le tableau 5.1, la variable aléatoire $\xi^{(3)}$ qui figure la qualité des conditions de logement prévoit quatre modalités (niveaux de qualité): « mauvais », « satisfaisant », « bon » et « très bon ». En attribuant à chaque famille sondée (d'après les critères adoptés) une de ces quatre modalités, on peut classer les familles par rapport à ce caractère. Le nombre total de modalités d'une variable ordinale peut être inférieur, égal et même supérieur au nombre d'objets sondés (d'expériences aléatoires).

La *variable aléatoire nominale* permet de diviser les objets sondés en classes non ordonnables homogènes pour le caractère analysé. Si l'analyste connaît le caractère analysé, ses modalités possibles et la règle d'attribution de ces modalités, alors la variable aléatoire ordinale ou nominale est dite *catégorisée*. Telles sont les variables aléatoires $\xi^{(1)}$ (milieu social de la famille) et $\xi^{(2)}$ (profession du chef de famille) du tableau 5.1. Si les conditions de l'expérience sont telles que son issue élémentaire est le comparateur δ_{ij} qui définit la mesure de ressemblance (ou de dissemblance), ou une relation d'ordre par rapport au caractère analysé, de deux objets i et j de la population sondée, alors une telle variable nominale est dite *non catégorisée*, et ses valeurs observées sont représentées sous la forme d'une *matrice de connexion* *)

$$\Delta = (\delta_{ij}), \quad i, j = 1, \ldots, n.$$

Un exemple de variable nominale non catégorisée est fourni par la variable décrivant une expérience aléatoire portant sur divers couples de familles afin de les classer par rapport à leur comportement de consommateur (cf. par exemple [137]).

Les problèmes de traitement statistique liés à l'étude d'un *caractère mixte* des variables aléatoires vectorielles, c'est-à-dire

*) Dans le cas particulier le plus fréquent, les éléments δ_{ij} ne peuvent prendre que deux valeurs: la valeur 1 si les objets i et j se voient attribuer la même modalité (ou sont rangés dans la même classe) et la valeur 0, sinon.

aux cas où comme dans le tableau 5.1, parmi les composantes figurent *simultanément* des composantes quantitatives, ordinales et nominales, sont abordés avec plus de détails au §§ 10.2 et 10.5.

5.4. Loi de probabilité d'une variable aléatoire. Population générale et échantillon

5.4.1. Loi de probabilité. Nous savons déjà (cf. n° 4.1.3) que pour décrire complètement le mécanisme de l'expérience aléatoire étudiée, c'est-à-dire l'espace probabilisé (ou ce qui revient au même la *variable aléatoire qui nous préoccupe*), *il ne suffit pas de définir seulement l'ensemble* Ω *des événements élémentaires* (*et, partant, décrire l'ensemble des valeurs théoriquement possibles de la variable aléatoire analysée*). A cela il faut ajouter: *pour le cas discret* — la loi qui associe à chaque valeur éventuelle X_i^0 de la variable aléatoire ξ la probabilité $p_i = P\{\xi = X_i^0\} = P\{\omega_i\}$ de son apparition; *pour le cas continu* — la loi qui associe à chaque domaine mesurable *) ΔX des valeurs éventuelles de la variable aléatoire ξ la probabilité $p(\Delta X) = P\{\xi \in \Delta X\}$ de l'événement « une valeur du domaine ΔX est réalisée au cours de l'expérience aléatoire ». Cette loi, qui est de la forme

$$\begin{aligned} \omega_i &\to p_i = P\{\xi = X_i^0\}; \\ \Delta X &\to p(\Delta X) = P\{\xi \in \Delta X\}, \end{aligned} \tag{5.2}$$

s'appelle *loi de probabilité* ou *distribution de probabilité* de la variable aléatoire ξ. La signification de ce terme est claire dans un espace probabilisé discret, car il est question de la *distribution de la probabilité 1* (c'est-à-dire la probabilité d'un événement certain) entre les diverses éventualités X_i^0 ($i = 1, 2, \ldots$).

Il est évident que la loi de probabilité (5.2) peut être donnée sous forme tabulaire et graphique (seulement dans le cas discret) ainsi que par des fonctions et des algorithmes (les principales formes de définition des lois de probabilité et les exemples de leurs notations analytiques sont exposés au chapitre 6).

Citons quelques exemples de lois de probabilité définies par des tableaux et des graphiques.

Une analyse statistique minutieuse de la pollution d'un échantillon par les pièces défectueuses (exemple 4.5) a permis de construire la distribution suivante des probabilités d'une variable aléatoire ξ exprimant le nombre de pièces défectueuses trouvées dans un échan-

---

*) Le domaine ΔX s'appelle *mesurable* si les issues élémentaires ω, correspondant aux valeurs de ce domaine, forment un *sous-ensemble mesurable* ou un *événement*, c'est-à-dire un sous-ensemble appartenant au système C de tous les événements possibles (cf. n°s 4.2.1 et 4.2.2).

tillon de taille $N = 30$ prélevé au hasard dans une population produite en série (tableau 5.2):

Tableau 5.2

i	1	2	3	4	5	6	7	8	9	10	11	12	...	31
x_i^0	0	1	2	3	4	5	6	7	8	9	10	11	...	30
$1000 \cdot p_i$	42	141	228	236	177	102	47	18	6	1	0	0	...	0

Les valeurs des probabilités du tableau 5.2 sont données au millième près, c'est pourquoi le fait que leur somme soit égale à 0,998 (au lieu de 1) admet une explication simple: le défaut 0,002 est « réparti » entre les valeurs 10, 11, . . ., 30 avec une probabilité inférieure à 0,0005 pour chaque valeur.

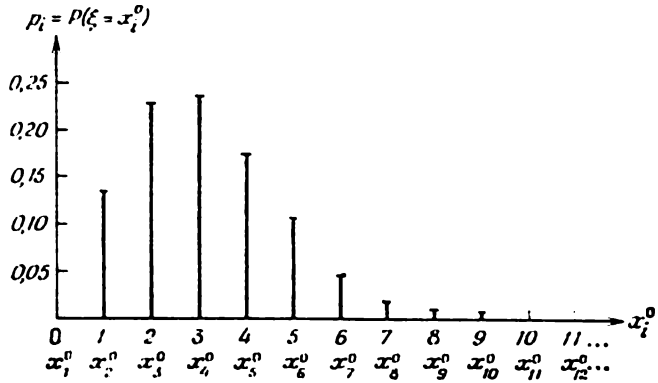


Fig. 5.2. Représentation graphique de la loi de probabilité du nombre de pièces défectueuses, découvertes dans un lot de 30 unités tirées au hasard dans une production en série

Cette loi de probabilité peut être représentée graphiquement (fig. 5.2).

La représentation géométrique de la loi de probabilité d'une variable aléatoire discrète s'appelle souvent *polygone de distribution* ou *polygone de fréquence.*

Considérons maintenant une partie du tableau 5.1 en ne retenant que deux des onze composantes: la qualité des conditions de logement $\xi^{(3)}$ et le revenu moyen par personne $\xi^{(6)}$. Simplifions davantage le schéma en passant de la variable continue $\xi^{(6)}$ à son analogue discret $\tilde{\xi}^{(6)}$ en renonçant à la connaissance exacte du revenu moyen par personne dans chaque famille et en se limitant à trois modalités:

la famille a un revenu bas (modalité $\tilde{x}_1^{(6)0}$), un revenu moyen (modalité $\tilde{x}_2^{(6)0}$) et un revenu élevé (modalité $\tilde{x}_3^{(6)0}$). En tenant compte des modalités attribuées à la qualité des conditions de logement : $x_1^{(3)0}$ pour mauvais ; $x_2^{(3)0}$ pour satisfaisant ; $x_3^{(3)0}$ pour bon et $x_4^{(3)0}$ pour très bon, et de l'analyse statistique effectuée, on obtient la loi de probabilité suivante pour la variable aléatoire bidimensionnelle $(\tilde{\xi}^{(6)}, \xi^{(3)})$ (les données sont arbitraires) :

Tableau 5.3

i \ j	$\tilde{x}_i^{(6)0}$ \ $x_j^{(3)0}$	1	2	3	4	$p_{i\cdot}$
		$x_1^{(3)0}$	$x_2^{(3)0}$	$x_3^{(3)0}$	$x_4^{(3)0}$	
1	$\tilde{x}_1^{(6)0}$	0,06	0,03	0,01	0,00	0,10
2	$\tilde{x}_2^{(6)0}$	0,05	0,25	0,35	0,05	0,70
3	$\tilde{x}_3^{(6)0}$	0,01	0,02	0,07	0,10	0,20
	$p_{\cdot j}$	0,12	0,30	0,43	0,15	1,00

Le polygone de fréquence correspondant est représenté sur la figure 5.3.

La loi de probabilité d'une variable aléatoire multidimensionnelle s'appelle *multidimensionnelle* ou *conjointe*. Si chaque composante $\xi^{(k)}$ ($k = 1, 2, \ldots, p$; cf. (5.1)) de la variable ξ analysée est discrète et prend un nombre m_k fini de valeurs, il est évident que le nombre total de « valeurs » du vecteur aléatoire ξ sera égal à $m = m_1 \cdot m_2 \cdot \ldots \cdot m_p$. Dans ce cas, au lieu de l'*indexation générale* de toutes les valeurs X_i^0 ($i = 1, 2, \ldots, m$) il est plus commode d'utiliser une *indexation p-dimensionnelle* de la forme $ij \ldots q$, où le premier indice i représente la i-ième valeur éventuelle de la première composante, le second indice j, la j-ième valeur éventuelle de la deuxième composante, etc. Alors $X_{ij\ldots q}^0$ désignera la valeur éventuelle de ξ obtenue en combinant la i-ième valeur éventuelle de la composante $\xi^{(1)}$, la j-ième valeur éventuelle de la composante $\xi^{(2)}$, ..., la q-ième valeur éventuelle de la composante $\xi^{(p)}$, et de désigner les

probabilités $P_1^T\{\xi = X^0_{ij\ldots q}\}$ par $p_{ij\ldots q}$. Donc, le tableau 5.3 représente les probabilités $p_{ij} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)^0},\ \xi^{(3)} = x_j^{(3)^0}\}$.

Souvent il est nécessaire de connaître la loi de probabilité d'*une partie seulement des composantes* de la variable aléatoire vectorielle analysée. C'est ainsi que la variable ξ du tableau 5.1 se décompose naturellement en deux sous-vecteurs: $\xi_1 = (\xi^{(1)}, \ldots, \xi^{(6)})$ et $\xi_2 = (\xi^{(7)}, \ldots, \xi^{(11)})$ décrivant respectivement les caractères sociodémographiques et la structure de la consommation de la famille.

La **loi de probabilité marginale** d'un sous-vecteur ξ_1 d'une variable aléatoire vectorielle $\xi = (\xi_1, \xi_2)$ décrit la répartition des probabilités de ξ_1 dans le cas où les valeurs de ξ_2 ne sont astreintes à aucune condition. Dans le cas discret, ces probabilités sont définies par les formules:

$$p_{i\cdot} = P\{\xi_1 = X_i^{(1)^0}\} = \sum_j P\{\xi_1 = X_i^{(1)^0};\ \xi_2 = X_j^{(2)^0}\}; \tag{5.3}$$

$$p_{\cdot j} = P\{\xi_2 = X_j^{(2)^0}\} = \sum_i P\{\xi_1 = X_i^{(1)^0},\ \xi_2 = X_j^{(2)^0}\}, \tag{5.3'}$$

où $X_i^{(1)^0}$ et $X_j^{(2)^0}$ représentent les i-ième et j-ième valeurs éventuelles des variables vectorielles ξ_1 et ξ_2 respectivement.

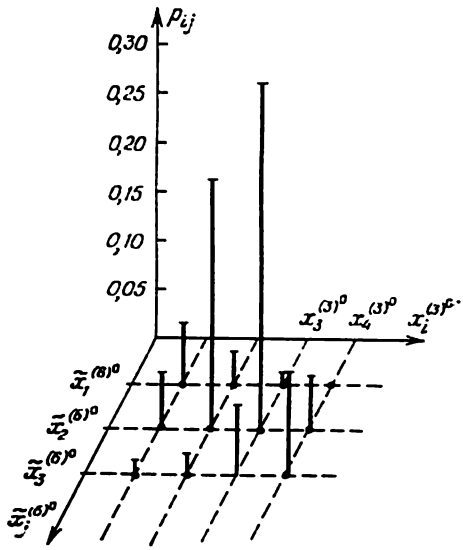


Fig. 5.3. Polygone de fréquence de la loi bidimensionnelle des familles suivant les conditions de logement ($\xi^{(3)}$) et le niveau des revenus ($\tilde{\xi}^{(6)}$)

$(p_{ij} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)^0},\ \xi^{(3)} = x_j^{(3)^0}\})$

Les formules (5.3) et (5.3′) résultent directement du théorème d'addition des probabilités (4.9″) si l'on tient compte des relations évidentes suivantes entre les événements considérés:

$$\{\xi_1 = X_i^{(1)^0}\} = \{\xi_1 = X_i^{(1)^0};\ \xi_2 = X_1^{(2)^0}\} + \{\xi_1 = X_i^{(1)^0};\ \xi_2 = X_2^{(2)^0}\} + \ldots;$$

$$\{\xi_2 = X_j^{(2)^0}\} = \{\xi_1 = X_1^{(1)^0};\ \xi_2 = X_j^{(2)^0}\} + \{\xi_1 = X_2^{(1)^0};\ \xi_2 = X_j^{(2)^0}\} + \ldots$$

Dans l'exemple envisagé (cf. tableau 5.3) les probabilités marginales $p_{i\cdot} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)^0}\}$ et $p_{\cdot j} = P\{\xi^{(3)} = x_j^{(3)^0}\}$ sont calculées à l'aide des formules (5.3) et (5.3′) et définissent respectivement la

répartition des familles séparément d'après leurs conditions de logement et d'après leur revenu (ces probabilités sont citées respectivement dans la dernière ligne et la dernière colonne du tableau 5.3).

La **loi de probabilité conditionnelle** du sous-vecteur ξ_1 d'une variable aléatoire vectorielle $\xi = (\xi_1, \xi_2)$ sachant que le sous-vecteur ξ_2 prend la valeur $X_j^{(2)\circ}$ est donnée par la formule

$$p_i(X_j^{(2)\circ}) = P\{\xi_1 = X_i^{(1)\circ} \mid \xi_2 = X_j^{(2)\circ}\} = \frac{P\{\xi_1 = X_i^{(1)\circ};\ \xi_2 = X_j^{(2)\circ}\}}{P\{\xi_2 = X_j^{(2)\circ}\}} = \frac{p_{ij}}{p_{\cdot j}}. \qquad (5.4)$$

De façon analogue,

$$p_j(X_i^{(1)\circ}) = P\{\xi_2 = X_j^{(2)\circ} \mid \xi_1 = X_i^{(1)\circ}\} = \frac{P\{\xi_1 = X_i^{(1)\circ};\ \xi_2 = X_j^{(2)\circ}\}}{P\{\xi_1 = X_i^{(1)\circ}\}} = \frac{p_{ij}}{p_{i\cdot}}. \qquad (5.4')$$

Les formules (5.4) et (5.4′) découlent directement du théorème de multiplication des probabilités (4.11).

Si, par exemple, l'on s'intéresse à la répartition conditionnelle d'un groupe de familles à revenu élevé par rapport à la qualité de leurs conditions de logement, c'est-à-dire à la répartition $p_{\cdot j}(\tilde{x}_3^{(6)\circ}) = P\{\xi^{(3)} = \tilde{x}_j^{(3)\circ} \mid \tilde{\xi}^{(6)} = \tilde{x}_3^{(6)\circ}\}$, alors les calculs effectués sur les données du tableau 5.3 à l'aide de la formule (5.4) nous conduisent aux résultats:

$$p_{\cdot 1}(\tilde{x}_3^{(6)\circ}) = \frac{0{,}01}{0{,}20} = 0{,}05;$$

$$p_{\cdot 2}(\tilde{x}_3^{(6)\circ}) = \frac{0{,}02}{0{,}20} = 0{,}10;$$

$$p_{\cdot 3}(\tilde{x}_3^{(6)\circ}) = \frac{0{,}07}{0{,}20} = 0{,}35;$$

$$p_{\cdot 4}(\tilde{x}_3^{(6)\circ}) = \frac{0{,}10}{0{,}20} = 0{,}50,$$

qui s'interprètent comme suit: parmi l'ensemble des familles *à revenu élevé*, 5 % vivent dans de mauvaises conditions de logement, 10 %, dans des conditions satisfaisantes, 35 %, dans de bonnes et 50 %, dans d'excellentes.

5.4.2. Population générale et échantillon. Ainsi, les lois qui régissent la variable aléatoire étudiée sont entièrement tributaires des conditions de son observation (ou de l'expérience) et sont définies *mathématiquement* par l'espace probabilisé correspondant $\{\Omega, C, P\}$ ou, ce qui revient au même, par la loi de probabilité correspondante.

Mais dans les analyses statistiques il semble plus commode d'opter pour une autre terminologie liée à la notion de population générale.

On appelle *population générale l'ensemble de toutes les observations possibles et imaginables (ou de tous les objets imaginables observés) susceptibles d'être réalisées sous les conditions données.* Etant donné que la définition parle de toutes les observations possibles et imaginables (ou de tous les objets), la notion de population générale est une notion mathématique *abstraite qu'il importe de ne pas confondre avec les populations réelles soumises à une étude statistique.* Ainsi, si l'on analyse même *toutes* les entreprises d'un secteur avec pour objectif l'enregistrement des valeurs caractérisant leurs indices techniques et économiques, on peut traiter la population sondée seulement comme le représentant d'une population hypothétique plus large d'entreprises susceptibles de fonctionner sous les mêmes conditions *).

Dans la pratique, il est plus commode de travailler sur les objets observés que sur leurs caractères. Nous prélevons pour les étudier des machines, des échantillons géologiques, des personnes et non les valeurs de leurs caractères. Par ailleurs, la théorie mathématique ne fait pas de distinction entre les objets et l'ensemble de leurs caractères et dissipe l'ambiguïté introduite par la définition.

Nous voyons que la notion mathématique de « population générale » de même que les notions d'« espace probabilisé », de « variable aléatoire » et de « loi de probabilité » dépendent entièrement des conditions d'observation (ou de l'expérience), donc on peut dans un certain sens considérer ces notions mathématiques comme *synonymes.* Une population générale est dite *finie* ou *infinie* selon que celle de toutes les observations imaginables est finie ou infinie.

De cette définition il résulte que les populations générales continues (composées d'observations de variables continues) sont toujours infinies. Les populations générales discrètes peuvent, elles, être aussi bien finies qu'infinies. Considérons l'exemple suivant (cf. n° 4.1.3): on analyse un lot de N pièces à la qualité; les pièces sont réparties en pièces de 1-re, 2-ème, 3-ème et 4-ème qualité; la variable aléatoire ξ est le numéro de la qualité d'une pièce prélevée au hasard; l'ensemble des valeurs éventuelles de la variable aléatoire ξ est composé de quatre points (1, 2, 3 et 4). Il est évident que nous avons affaire ici à une population générale finie (au total, N observations imaginables).

La notion de population générale infinie est une abstraction mathématique au même titre que la possibilité de répéter une infinité

*) Il importe de faire une distinction entre l'ensemble de toutes les observations imaginables et celui de toutes les valeurs imaginables (ou théoriquement possibles) de la variable aléatoire ξ_i; les observations sont généralement en plus grand nombre, puisque à toute valeur possible X peuvent correspondre plusieurs ou même un nombre infini d'observations imaginables.

de fois la mesure d'une variable aléatoire. La population générale infinie peut être traitée approximativement comme le cas limite d'une population finie dont le nombre des objets étudiés sous des conditions données croît indéfiniment. Si dans l'exemple ci-dessus on remplace le lot de N pièces par une production en série des mêmes pièces, on sera conduit à la notion de population générale infinie. Pratiquement cela revient à exiger que $N \to \infty$.

Un échantillon d'une population générale donnée est le résultat d'une suite finie d'observations $X_1, X_2, \ldots, X_n$ *d'une variable aléatoire* ξ. L'échantillon peut être regardé comme une sorte d'analogue empirique de la population générale, ce à quoi nous sommes le plus souvent confrontés, puisque le sondage de la population générale soit est trop compliqué (pour de grands N), soit est impossible (pour N infini). Le nombre n d'observations s'appelle *taille de l'échantillon.*

Si la taille n est grande ($n > 50$) et la variable aléatoire étudiée, continue et à une dimension (ou bien discrète, à une dimension et prenant un nombre de valeurs assez élevé, disons, plus de 10), il est souvent plus commode, dans la perspective d'une simplification du traitement statistique des résultats des observations, de passer aux données « groupées » d'échantillonnage. Ce passage s'effectue généralement de la manière suivante:

a) on relève la plus petite valeur, $x_{\min}(n)$, et la plus grande, $x_{\max}(n)$, dans l'échantillon;

b) on subdivise l'intervalle $[x_{\min}(n), x_{\max}(n)]$ qui s'appelle *étendue de l'échantillon* en un nombre s donné d'intervalles de groupement égaux; le nombre s ne doit pas être inférieur à $8 \sim 10$ et supérieur à $20 \sim 25$: le choix du nombre d'intervalles s dépend essentiellement de la taille n de l'échantillon; pour avoir une idée approximative de la valeur de s, on peut se servir de la formule

$$s \approx \log_2 n + 1$$

qu'il faut plutôt comprendre comme une formule définissant un minorant de s (surtout pour n grand);

c) on repère les bornes de chaque intervalle, c'est-à-dire les points $c_0, c_1, c_2, \ldots, c_s$ dans leur ordre de croissance, ainsi que les points milieux $x_1^0, x_2^0, \ldots, x_s^0$;

d) on fait le compte des données d'échantillonnage comprises dans chaque intervalle: il est évident que la somme $\nu_1 + \nu_2 + \ldots + \nu_s$ des nombres de ces données est égale à n: les données d'échantillonnage qui sont confondues avec les bornes des intervalles sont soit uniformément réparties sur les deux intervalles voisins respectifs, soit par convention rapportées à l'un d'eux, celui de gauche pour fixer les idées.

Cette procédure de groupement peut être modifiée en fonction de la teneur du problème abordé (dans certains cas, par exemple, on aura intérêt à prendre des intervalles de groupement inégaux).

On adoptera les notations ci-dessus dans tous les raisonnements ultérieurs faisant intervenir des données d'échantillonnage.

On rappelle que les méthodes statistiques sont destinées à analyser les caractères d'une population sur un de ses échantillons.

Un problème important dont la solution conditionne la crédibilité des résultats fournis par le traitement statistique des données initiales est le *problème de représentativité de l'échantillon*, c'est-à-dire le problème de savoir dans quel degré sont présents dans l'échantillon les caractères de la population générale auxquels on s'intéresse. Un même groupe d'objets peut être interprété comme un échantillon prélevé dans des populations différentes. Ainsi, un groupe de familles vivant dans des copropriétés rattachées à un Office du Logement d'un arrondissement peut être, dans le cadre d'un sondage sociologique, traité soit comme un échantillon de la population générale des familles vivant dans des copropriétés gérées par un Office de Logement donné, soit comme un échantillon de la population générale des familles d'un arrondissement donné, soit comme un échantillon de la population générale des familles de la ville, soit enfin comme un échantillon de la population générale des familles de la ville habitant des copropriétés. Les résultats du sondage s'interprètent différemment selon que le groupe de familles prélevé fait partie de telle ou telle population générale, selon aussi la population générale pour laquelle il est représentatif. La réponse à ce problème est liée à plusieurs facteurs. Notamment dans l'exemple ci-dessus à la présence-absence d'un facteur spécial (éventuellement latent) définissant le rattachement de la famille à un Office de Logement donné ou à un arrondissement (ce facteur pourrait être le revenu moyen de la famille, la position géographique de l'arrondissement dans la ville, l'« âge » de l'arrondissement, etc.).

5.4.3. Principales méthodes d'organisation de l'échantillonnage. Dans l'estimation de la représentativité d'un échantillon entrent en ligne de compte et le procédé de prélèvement de l'échantillon et le degré avec lequel la répartition des variables principales analysées dans l'échantillon est caractéristique pour la population générale. Le premier moyen d'accroître la représentativité, c'est-à-dire de réaliser un tirage parfaitement aléatoire, est souvent difficile sur le plan organisationnel. Par ailleurs, la combinaison d'un tirage aléatoire et d'un tirage régulier est parfois plus payante. En tout cas, la méthode de collecte des données initiales doit être minutieusement élaborée et entièrement décrite dans le compte rendu.

L'utilisation des répartitions des principaux critères dans l'échantillon et dans la population générale pour estimer la représentativité présente ses propres difficultés dont certaines revêtent un caractère typiquement statistique: petite taille de l'échantillon, mauvaise élaboration des méthodes de comparaison des distributions conjointes,

etc., d'autres, un caractère organisationnel : en effet, on n'est pas toujours fixé *a priori* sur les variables dont il faut comparer les répartitions pour démontrer la représentativité.

Décrivons maintenant les principales méthodes d'organisation de l'échantillonnage.

Tirage aléatoire simple. Ce procédé consiste à tirer n objets d'une population générale de taille N. On obtient ainsi C_N^n échantillons équiprobables. En pratique, les objets de la population générale sont numérotés de 1 à N, puis les objets de l'échantillon sont prélevés l'un après l'autre à l'aide d'une table de nombres aléatoires ou de tout autre méthode de tirage équiprobable (par exemple, une urne de N boules numérotées de 1 à N). Ce tirage est dit *aléatoire* ou *probabiliste.*

Le *tirage simple à l'aide d'une procédure régulière sans conséquence sur le problème étudié* remplace souvent le tirage aléatoire. En médecine, le tirage peut être effectué d'après le jour de la semaine, ce qui est commode sur le plan de l'organisation ; dans les sondages sociologiques, d'après les initiales des noms de famille des personnes rattachées à un Office de Logement donné, etc. Ces tirages sont dits *mécaniques.*

La *stratification* consiste à subdiviser la population générale initiale de taille N en sous-populations de taille $N_1, N_2, \ldots, N_R$, deux à deux disjointes, c'est-à-dire telles que $N_1 + N_2 + \ldots + N_R = N$. Les sous-populations s'appellent *strates.* Lorsque les strates sont définies, on extrait de chacune d'elles un échantillon aléatoire simple respectivement de taille $n_1, n_2, \ldots, n_R$. Pour tirer entièrement parti des strates, il faut connaître les valeurs de N_i/N. Le tirage stratifié est utilisé lorsque les *strates sont homogènes* en ce sens que les objets qui les composent présentent des caractères voisins (dont les valeurs moyennes peuvent être déterminées sur de petits échantillons) ; lorsque la population générale doit être analysée avec la même minutie à travers toutes ses strates ; lorsque, pour des raisons d'organisation, les méthodes de tirage dans les strates sont différentes. Ces échantillons sont dits *stratifiés.*

Un cas particulier de tirage stratifié est la méthode qui consiste à sélectionner les strates d'après un critère *indirect* lié d'une certaine manière au critère étudié. Ainsi, pour obtenir un échantillon stratifié lors de l'étude du revenu moyen par personne dans une famille, on peut préalablement diviser la population de familles en groupes homogènes par rapport à un quelconque caractère socio-économique du chef de famille (par exemple, le salaire). Dans ces cas, on parle de méthode de tirage *typique* et respectivement d'échantillons *typiques.*

Les *méthodes de tirage par grappes* sont mises en œuvre lorsqu'il est plus commode de travailler non pas sur des éléments isolés, mais sur plusieurs, liés entre eux dans le temps ou dans l'espace.

Cet ensemble forme une « grappe » et les éléments « les graines » de la grappe. Ainsi, le mode d'administration par régions et le caractère de la gestion administrative plaident en faveur d'un sondage continu d'unités régionales (immeubles, quartiers) et non pas de familles isolées.

L'*échantillonnage combiné à plusieurs degrés* consiste à distinguer en cascade plusieurs degrés et à appliquer à chacun d'eux une méthode de tirage. Si par exemple l'on étudie les conditions de vie et la structure des familles d'une ville, le premier degré consistera à définir les arrondissements par un tirage aléatoire, le deuxième, à désigner les Offices de logement par un tirage mécanique, le troisième, à sélectionner les immeubles par un tirage par grappes.

La théorie et les méthodes de sondage sont décrites en détail, par exemple, dans [36].

Tirage séquentiel (actif). Quand on analyse des processus physico-chimiques et technologiques, on est souvent amené à étudier des relations de la forme

$$y = f(X, \Theta) + \xi,$$

où f est une fonction connue, $X = (x^{(1)}, \ldots, x^{(p)})'$ un ensemble de variables de commande, Θ un point inconnu de l'espace des paramètres (« l'état de la nature »), ξ une erreur aléatoire de moyenne nulle. On demande d'apprécier Θ en un nombre minimal d'expériences aux points $X_1, X_2, \ldots, X_n$ du domaine R^p des valeurs admissibles de X. Les expériences peuvent être effectuées en tout point X du domaine R^p. Dans ces conditions, après chaque série d'expériences, on traite les données recueillies pour déterminer la série suivante de points X qui contient le plus d'information sur Θ ; on répète les expériences en ces points pour trouver une nouvelle estimation de Θ et une nouvelle série de points X, et ainsi de suite. Pour spécifier que les points X ne sont pas choisis au petit bonheur, ces expériences sont dites *actives*. Pour plus de détails sur les problèmes de ce type, voir, par exemple, [52].

5.5. Méthodes de définition de la loi de probabilité : fonction de répartition, densité de probabilité et leurs analogues empiriques

5.5.1. Fonction de répartition d'une variable aléatoire à une dimension. On a vu plus haut (cf. § 5.4) que toute population générale (variable aléatoire) est définie par sa loi de probabilité $p(\Delta X)$. Etant donné que les domaines ΔX peuvent être généralement des sous-ensembles de même nature, la question suivante se pose : quelles sont les méthodes de définition des fonctions numériques p sur ΔX qui sont assez commodes sur le plan pratique ?

Il s'avère que pour décrire la répartition de variables aléatoires ξ

à une dimension, il suffit de se donner un procédé de calcul des probabilités $p(\Delta X) = P\{\xi \in \Delta X\}$ uniquement pour des sous-ensembles ΔX d'un type spécial, plus exactement, pour des intervalles semi-ouverts à droite:

$$\Delta X = [x_{\min}, x[$$

où $x_{\min}$ est la plus petite valeur de la variable aléatoire ξ (cette valeur peut éventuellement être égale à $-\infty$), x une valeur « courante » (c'est-à-dire définie par nous) possible de ξ. La probabilité $p([x_{\min}, x[) = P\{\xi < x\}$ est définie de façon unique par la donnée de l'extrémité droite de l'intervalle, c'est-à-dire par le *nombre* x, et peut donc être interprétée comme une ordinaire fonction de x.

La *fonction de répartition des probabilités* (ou *fréquence cumulée*) $F_\xi(x)$ d'une variable aléatoire ξ est une fonction qui associe à toute valeur donnée x la valeur de la probabilité de l'événement $\{\xi < x\}$, soit

$$F_\xi(x) = P\{\xi < x\} \tag{5.5}$$

Dans la suite, si aucune confusion n'est à craindre, on omettra l'indice ξ de la fonction F et on l'appellera simplement *fonction de répartition*.

Etudions le comportement de la fonction de répartition. Signalons tout d'abord que dans le cas *discret* l'événement $A(x) = \{\xi < x\}$ est composé de tous les événements élémentaires $\omega_i = \{\xi = x_i^0\}$ tels que $x_i^0 < x$. En vertu de la définition de la probabilité d'un événement composé (cf. n° 4.1.3), on obtient

$$F(x) = P\{\xi < x\} = \sum_{i:\, x_i^0 < x} P\{\xi = x_i^0\} = \sum_{i:\, x_i^0 < x} p_i \tag{5.5'}$$

(la sommation est effectuée sur tous les i, tels que $x_i^0 < x$).

Sur (5.5') on voit que la fonction $F(x)$ est *en escalier*, la valeur de chaque saut étant égale à $p_i = P\{\xi = x_i^0\}$.

La situation est différente lorsque la variable aléatoire ξ est *continue*. La plupart des variables aléatoires continues, dont l'étude présente un intérêt pratique, sont telles que pour tout intervalle Δx les probabilités $P\{\xi \in \Delta x\}$ tendent vers 0 lorsque Δx tend vers 0, donc, les *probabilités que ξ prenne des valeurs isolées x* sont nulles (un exemple analogue est cité au n° 4.2.2 dans le problème du sondage des experts). Il est aisé de comprendre que les fonctions de répartition de telles variables aléatoires sont *continues*. La figure 5.4 représente les graphiques des fonctions de répartition des variables aléatoires envisagées dans les exemples 4.1, 4.2, 4.5 (compte tenu du tableau 5.2) et dans l'exemple du sondage des experts (cf. n° 4.2.2).

De la définition de la fonction de répartition on déduit immédiatement les propriétés suivantes.

a) $F_\xi(x)$ est une fonction croissante de x;
b) $F_\xi(x) = 0$ pour tous les $x \leqslant x_{\min}$;
c) $F_\xi(x) = 1$ pour tous les $x > x_{\max}$ ($x_{\min}$ et $x_{\max}$ sont respectivement les valeurs minimale et maximale de la variable aléatoire ξ);
d) $P\{a \leqslant \xi < b\} = F_\xi(b) - F_\xi(a)$ quels que soient a et b (pour établir cette propriété, il faut utiliser le théorème d'addition

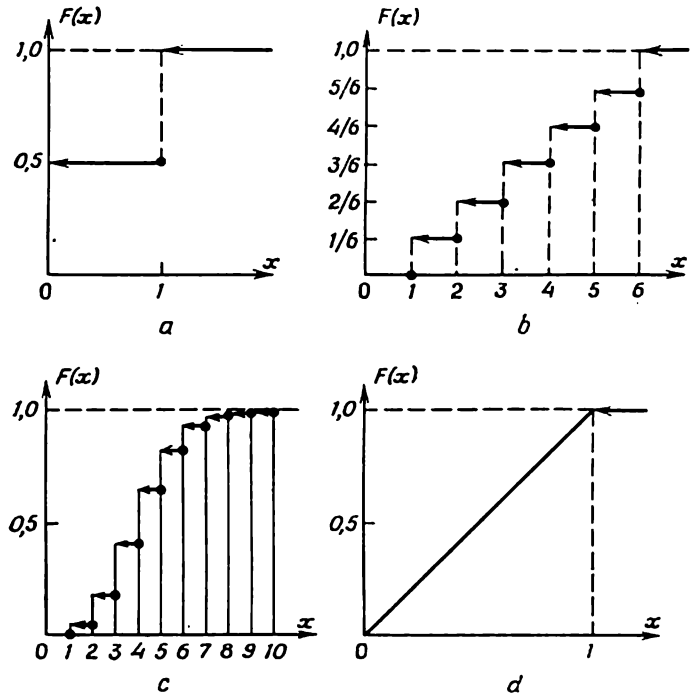


Fig. 5.4. Représentation graphique de la fonction de répartition pour: *a*, le résultat chiffré du jet d'une pièce de monnaie (le zéro correspond à pile, le un à face); *b*, nombre de points obtenus en lançant un dé symétrique; *c*, nombre de pièces défectueuses, découvertes dans un lot de 30 unités tirées au hasard (cf. table 5.2); *d*, estimation par des experts incompétents de la probabilité d'un événement donné, cf. exemples des n°s 2.1.3 et 4.2.2

des probabilités (cf. n° 4.1.3) et le fait que les événements $A = \{\xi < a\}$, $B = \{\xi < b\}$ et $C = \{a \leqslant \xi < b\}$ sont liés par la relation $B = A + C$).

La forme exacte de la fonction de répartition est généralement inconnue dans la pratique statistique. On appelle analogue *empirique* (ou *d'échantillonnage*, c'est-à-dire construit à l'aide d'un échantillon de taille n) de la fonction de répartition théorique $F(x)$ la fonc-

tion $\hat{F}^{(n)}(x)$ définie par les relations

$$\hat{F}^{(n)}(x) = \frac{\nu(x)}{n} \tag{5.6}$$

ou si les données sont groupées (cf. n° 5.4.2)

$$\hat{F}^{(n)}(x) = \frac{\nu_1 + \nu_2 + \ldots + \nu_{i_x}}{n}. \tag{5.6'}$$

où $\nu(x)$ est le nombre des valeurs de ξ strictement inférieures à x observées dans l'échantillon $x_1, x_2, \ldots, x_n$; ν_i le nombre des valeurs observées, appartenant au i-ème intervalle de groupement ; i_x le numéro du plus à droite des intervalles de groupement dont l'extrémité droite est inférieure à x. La définition de la fonction de répartition empirique permet d'expliquer l'appellation de « fréquence cumulée » qui est souvent donnée à la fonction de répartition. La propriété de stabilité statistique des fréquences relatives (cf. § 7.2) justifie l'emploi de $F^{(n)}(x)$ à titre d'approximation de la fonction de répartition théorique inconnue $F(x)$ et le fait que l'erreur de cette approximation tend vers 0 lorsque n tend vers l'infini. Une telle approximation de $F(x)$, c'est-à-dire une approximation qui n'est pas liée au choix préalable d'un modèle général de cette fonction *), est dite *non paramétrique*. Aux §§ 10.3 et 11.1 on trouvera de plus amples détails sur l'étude statistique des fonctions de répartition empiriques.

5.5.2. Densité de probabilité d'une variable aléatoire à une dimension. Les variables aléatoires continues, dont les fonctions de répartition sont partout continues et différentiables, peuvent être également définies par leur *densité de probabilité* $f_\xi(x)$:

$$f_\xi(x) = \lim_{\Delta \to 0} \frac{P\{\xi \in [x, x+\Delta[\}}{\Delta} = \lim_{\Delta \to 0} \frac{F_\xi(x+\Delta) - F_\xi(x)}{\Delta}, \tag{5.7}$$

ou, ce qui est équivalent,

$$f_\xi(x) = F'_\xi(x), \tag{5.7'}$$

autrement dit, $f_\xi(x)$ est la dérivée de la fonction de répartition $F_\xi(x)$ au point x. Des relations équivalentes (5.7) et (5.7') on déduit immédiatement les propriétés suivantes de la densité de probabilité $f_\xi(x)$:

a) $f_\xi(x) \geqslant 0$, puisque $F_\xi(x)$ est croissante ;

b) $P\{\xi \in [x, x+\Delta[\} \approx f_\xi(x) \cdot \Delta$ pour les petits Δ (ceci résulte de la comparaison des deux premiers membres de l'identité (5.7)) ;

c) $P\{\xi \in [x_{\min}, x[\} = F_\xi(x) = \int_{x_{\min}}^{x} f(u)\,du$ pour tout x ;

*) Le chapitre 6 traite des fonctions de répartition les plus fréquemment utilisées en statistique.

d) $P\{\xi \in [x_0, x_0+\Delta[\} = F_\xi(x_0+\Delta) - F_\xi(x_0) = \int\limits_{x_0}^{x_0+\Delta} f(u)\,du$ pour tous x_0 et Δ;

e) $P\{\xi \in [x_{\min}, x_{\max}]\} = \int\limits_{x_{\min}}^{x_{\max}} f(x)\,dx \equiv 1.$

Attardons-nous sur quelques propriétés de la densité de probabilité.

La propriété b) permet de clarifier le sens probabiliste de la densité de probabilité. Ainsi, supposons, pour fixer les idées, que le domaine $[x_{\min}, x_{\max}]$ des valeurs possibles de la variable aléatoire ξ est fini; subdivisons-le en intervalles de groupement Δ égaux et assez fins de centres

$$x_1^0 = x_{\min} + \frac{\Delta}{2},$$
$$x_2^0 = x_1^0 + \Delta,$$
$$x_3^0 = x_2^0 + \Delta, \ \ldots$$

et associons à chaque intervalle i la probabilité de réalisation de l'événement

$$\xi \in \left[x_i^0 - \frac{\Delta}{2}, x_i^0 + \frac{\Delta}{2}\right],$$

probabilité qui est égale approximativement à $f_\xi(x_i^0)\cdot\Delta$ d'après la propriété b). Donc, les valeurs de la fonction $f_\xi(x)$ sont proportionnelles à la probabilité que la variable aléatoire ξ prenne une valeur voisine de x. Ceci explique, en particulier, que l'*analogue discret de la densité de probabilité d'une variable aléatoire discrète est le polygone de fréquence, c'est-à-dire la suite des points de coordonnées* (x_i^0, p_i). De là, il s'ensuit notamment que la valeur *la plus fréquente* (dite encore *mode* ou *dominante*) de la variable aléatoire continue ξ est la valeur x_{mod} qui maximise la densité de probabilité, c'est-à-dire $f_\xi(x_{\text{mod}}) = \max\limits_x f_\xi(x)$.

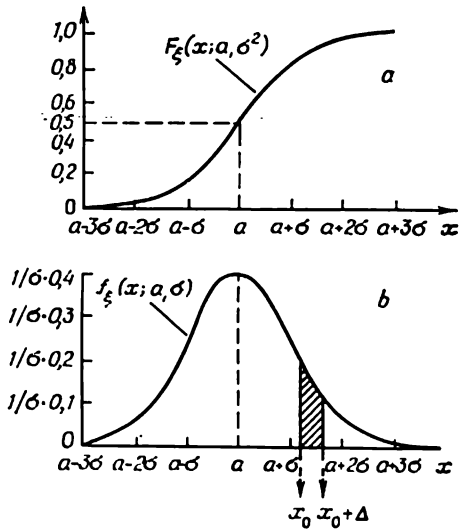


Fig. 5.5. (*a*) Fonction de répartition $F_{\xi_{\text{norm}}}(x; a, \sigma^2)$ et (*b*) densité de probabilité de la loi normale $f_{\xi_{\text{norm}}}(x; a, \sigma^2)$

La propriété d) admet l'interprétation géométrique suivante: la probabilité de l'événement $\{\xi \in [x_0, x_0 + \Delta[\}$ est (quels que soient

x_0 et Δ donnés) égale à l'aire du trapèze curviligne de base $[x_0, x_0 + \Delta[$ situé sous la courbe de la densité $y = f_\xi(x)$.

La figure 5.5 représente la fonction de répartition $F_\xi(x; a, \sigma^2)$ et la densité de probabilité $f_\xi(x; a, \sigma^2)$ de la loi *normale* (pour plus de détails, cf. §§ 6.1 et 7.3). L'aire hachurée de la figure 5.5, *b* représente la valeur de la probabilité $P\{\xi \in [x_0, x_0 + \Delta[\}$.

Les données empiriques $x_1, x_2, \ldots, x_n$ permettent de se faire une idée approximative de la densité de probabilité théorique inconnue $f_\xi(x)$. Si l'on s'intéresse à l'intervalle $[x, x + \Delta[$ et si l'on y a dénombré $n_\Delta(x) = \nu(x + \Delta) - \nu(x)$ observations, il est évident que l'analogue empirique de la quantité

$$\frac{F_\xi(x+\Delta) - F_\xi(x)}{\Delta} \approx f_\xi(x)$$

sera la quantité

$$\frac{\frac{\nu(x+\Delta)}{n} - \frac{\nu(x)}{n}}{\Delta} = \frac{n_\Delta(x)}{n \cdot \Delta} = \hat{f}_\xi^{(n)}(x). \tag{5.8}$$

Il est évident que la valeur de $\hat{f}_\xi^{(n)}(x)$ caractérise la *densité des observations* de la variable aléatoire ξ au voisinage du point x, c'est-à-dire la fréquence relative des observations, rapportée à une unité de longueur de l'intervalle des valeurs possibles de ξ. Pour cette raison, la fonction $\hat{f}_\xi^{(n)}(x)$ définie par (5.8) s'appelle *densité de probabilité empirique* (ou *d'échantillonnage*). Cette circonstance justifie le choix de cette terminologie pour la densité *théorique* $f_\xi(x)$ dans la mesure où, en vertu de la stabilité statistique des fréquences (cf. § 7.2), ces deux quantités se rapprochent indéfiniment lorsque $n \to \infty$ et $\Delta \to 0$.

Pour construire la densité de probabilité empirique $\hat{f}_\xi^{(n)}(x)$ sur son domaine tout entier de définition (c'est-à-dire *pour toutes* les valeurs possibles de la variable ξ), on se sert des *données groupées* (cf. n° 5.4.2) et l'on pose

$$\hat{f}_\xi^{(n)}(x) = \frac{\nu_{k(x)}}{n \cdot \Delta_{k(x)}}, \tag{5.8'}$$

où $k(x)$ est l'indice de l'intervalle de groupement contenant le point x; $\nu_{k(x)}$ et $\Delta_{k(x)}$ respectivement le nombre d'observations appartenant à cet intervalle et sa longueur. La représentation géométrique de la densité de probabilité empirique s'appelle *histogramme*.

Exemple 5.2. Objet du sondage: 995 abonnés au téléphone de la ville de Buffalo (État de New York). But du sondage: détermination du nombre ξ de communications passées par chaque abonné durant une année. Dans ce cas, donc, $n = 995$, x_i est le nombre de

communications de l'abonné i *) durant une année. Le tableau 5.4 représente les données groupées d'après la méthode décrite au n° 5.4.2. Le nombre d'intervalles de groupement a été pris égal à $k = 22$ et la longueur d'un intervalle Δ à 50. Les graphiques des densités de probabilité empirique (l'histogramme) et théorique sont représentés

Tableau 5.4

Numéro de l'intervalle de groupement	Nombre d'observations tombant dans l'intervalle envisagé	Numéro de l'intervalle de groupement	Nombre d'observations tombant dans l'intervalle envisagé
1	0	12	116
2	1	13	79
3	9	14	54
4	19	15	31
5	38	16	11
6	50	17	5
7	95	18	6
8	85	19	2
9	115	20	1
10	132	21	1
11	144	22	1

sur la figure 5.6 (pour construire approximativement la courbe représentative de la densité théorique, on a remplacé les paramètres inconnus a (la moyenne) et σ^2 (la variance) par les valeurs des caractéristiques empiriques correspondantes $\hat{a}$ et $\hat{\sigma^2}$ (cf. § 10.4)).

Les méthodes de construction des densités de probabilité empiriques et leur analyse statistique sont abordées avec plus de détails dans les §§ 10.3, 10.4, 11.1 et 11.2.

5.5.3. Fonction de répartition et densité de probabilité conjointes. Indépendance statistique des variables aléatoires. Ce qui précède met l'accent sur la nécessité d'élaborer des méthodes commodes de définition de la loi de probabilité d'une variable aléatoire continue: la forme universelle de description du « comportement » d'une variable aléatoire discrète ξ (à condition que le nombre de ses valeurs possibles ne soit pas trop élevé) est le *polygone de fréquence* qui à toute valeur possible x_i^0 de ξ associe la probabilité de sa réalisation $p_i = P\{\xi = X_i^0\}$.

*) Cf. *Introduction to Frequency Curves.* — Bull. Amer. tel. and telegr. Company, 1953, n° 1. Ceci est précisément le cas où pour le traitement statistique d'une variable aléatoire *discrète* (le nombre de communications passées en une année en est indiscutablement une) il est commode d'utiliser la technique d'analyse statistique des variables aléatoires *continues* (passage aux données groupées, construction de la densité empirique, etc.), ce qui a été déjà signalé au § 5.2.

Nous allons donc porter notre attention sur le cas continu. Signalons qu'*à l'inverse de la dimension un, la fonction de répartition multidimensionnelle*

$$F_{(\xi^{(1)}, \ldots, \xi^{(p)})}(x^{(1)}, \ldots, x^{(p)}) = P\{\xi^{(1)} < x^{(1)}, \ldots, \xi^{(p)} < x^{(p)}\} \quad (5.9)$$

cesse d'être une forme commode de définition de la loi de probabilité étudiée.

Les analogues multidimensionnels des intervalles finis et semi-infinis (qui s'obtiennent par la somme et l'intersection d'intervalles semi-infinis de la forme $[-\infty, x[$) sont les *hyperparallélépipèdes* finis et semi-infinis. La fonction de répartition (5.9) définit la règle de calcul des probabilités pour des domaines multidimensionnels de ce type. Mais, si en dimension un cela suffisait pour « travailler » dans l'espace probabilisé correspondant, en dimension n la situation est différente. En particulier, la connaissance de la seule forme (5.9) ne suffit pas à résoudre le problème, très important pour les applications statistiques, qu'est la description de la loi de probabilité de fonctions des variables aléatoires initiales $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ (l'approche générale de la résolution de ce problème est décrite au § 7.4).

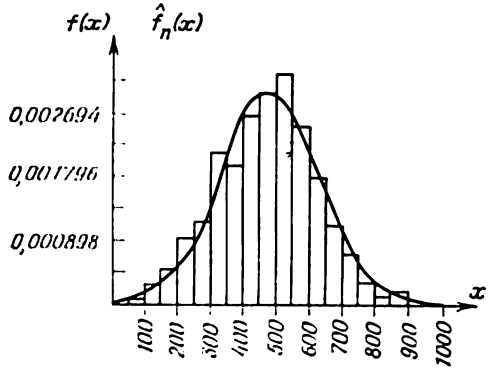


Fig. 5.6. Histogramme de $\hat{f}_n(x)$ et densité de probabilité $f(x)$ d'une loi normale convenablement choisie, caractérisant la répartition du nombre d'appels téléphoniques passés par un abonné en une année

Donc, pour décrire la loi de probabilité de la variable aléatoire continue $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$, on se sert de la densité de probabilité $f_\xi(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ qui peut être définie soit à l'aide de la fonction de répartition (5.9) comme suit:

$$f_\xi(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \frac{\partial^p F(x^{(1)}, \ldots, x^{(p)})}{\partial x^{(1)} \partial x^{(2)} \ldots \partial x^{(p)}}, \quad (5.10)$$

soit indépendamment d'elle *) comme suit: la densité de probabilité de la variable aléatoire $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ est une fonction

*) De même qu'en dimension un, cette description vaut uniquement pour les variables aléatoires continues dont la fonction de répartition (5.9) est continue par rapport à toutes les variables à l'intérieur du domaine des valeurs possibles et admet dans ce domaine une dérivée partielle continue par rapport à toutes ses variables (on dit que $F_\xi(x^{(1)}, \ldots, x^{(p)})$ est absolument continue). Comme déjà signalé, ces variables aléatoires constituent l'écrasante majorité des variables présentant un intérêt pratique.

$f(x^{(1)}, \ldots, x^{(p)})$ de p variables telle que, pour tout sous-ensemble (mesurable) A des valeurs possibles de ξ, la probabilité de l'événement $\{\xi \in A\}$ est donnée par la relation

$$P\{\xi \in A\} = \int_A f_\xi(x^{(1)}, \ldots, x^{(p)})\, dx^{(1)} \ldots dx^{(p)}, \qquad (5.10')$$

où l'intégration est étendue au domaine A donné de l'espace des valeurs possibles de ξ (c'est-à-dire que l'intégration dans (5.10') est une p-uple intégration).

La densité de probabilité admet la même interprétation qu'en dimension un: la probabilité de réalisation d'une valeur de ξ, comprise dans un petit voisinage $\Delta X = \{x^{(1)} \leqslant \xi^{(1)} < x^{(1)} + \Delta x^{(1)}, \quad x^{(2)} \leqslant \xi^{(2)} < x^{(2)} + \Delta x^{(2)}, \quad \ldots, \quad x^{(p)} \leqslant \xi^{(p)} < x^{(p)} + \Delta x^{(p)}\}$ du point $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, est proportionnelle à la valeur $f_\xi(X)$ de la densité de probabilité en ce point et est égale, en particulier, à l'« élément de probabilité » $f_\xi(X)\,\Delta X$, c'est-à-dire que

$$P\{x^{(1)} \leqslant \xi^{(1)} < x^{(1)} + \Delta x^{(1)}, \ldots, x^{(p)} \leqslant \xi^{(p)} < x^{(p)} + \Delta x^{(p)}\} \approx$$
$$\approx f_\xi(x^{(1)}, \ldots, x^{(p)})\, \Delta x^{(1)} \ldots \Delta x^{(p)}. \qquad (5.11)$$

Les *analogues empiriques* $\hat{F}_\xi^{(n)}(X)$ et $f_\xi^{(n)}(X)$ de la fonction de répartition $F_\xi(X)$ et de la densité de probabilité $f_\xi(X)$ théoriques se construisent à l'aide des données empiriques $X_1, X_2, \ldots, X_n$ de la variable aléatoire ξ comme en dimension un, soit

$$\hat{F}_\xi^{(n)}(X) = \frac{\nu^{(n)}(X)}{n}; \qquad (5.12)$$

$$\hat{f}_\xi^{(n)}(X) = \frac{\nu_k(X)}{n \cdot \Delta x^{(1)} \cdot \Delta x^{(2)} \cdot \ldots \cdot \Delta x^{(p)}}, \qquad (5.13)$$

où $X = (x^{(1)}, \ldots, x^{(p)})$ est la valeur qui nous intéresse de la variable ξ; $\nu^{(n)}(X)$ est le nombre des données empiriques X_i dont les composantes vérifient simultanément les conditions

$$\begin{cases} x_i^{(1)} < x^{(1)}; \\ x_i^{(2)} < x^{(2)}; \\ \cdots\cdots \\ x_i^{(p)} < x^{(p)}, \end{cases}$$

$k(X)$, le numéro de l'hyperparallélépipède de groupement contenant le point X, ν_k le nombre des données empiriques tombant dans le k-ième hyperparallélépipède de groupement *).

*) L'autre méthode de construction des densités empiriques, méthode qui n'implique pas de passer aux données groupées, est accessible au n° 10.3.2. Le passage aux données groupées en dimension n s'effectue comme en dimension un (cf. n° 5.4.2). Mais, pour que ce passage *ait un sens*, il faut évidemment tra-

Les fonctions définies par les relations (5.9), (5.10) ou (5.10′) s'appellent respectivement fonction de répartition *conjointe* et densité de probabilité *conjointe* de la variable aléatoire vectorielle $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$.

Pour décrire la loi de probabilité *marginale d'une partie* des composantes $\xi_1 = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(s)})$, $s < p$, du vecteur ξ (cf. n° 5.4.1) on se sert de la *fonction de répartition marginale* $F_{\xi_1}(x^{(1)}, \ldots, x^{(s)})$ et de la *densité de probabilité marginale*, définies respectivement par les formules :

$$F_{\xi_1}(x^{(1)}, \ldots, x^{(s)}) = P\{\xi^{(1)} < x^{(1)}, \ldots, \xi^{(s)} < x^{(s)}\} =$$
$$= P\{\xi^{(1)} < x^{(1)}, \ldots, \xi^{(s)} < x^{(s)}, \xi^{(s+1)} < \infty, \ldots, \xi^{(p)} < \infty\} =$$
$$= F_\xi(x^{(1)}, x^{(2)}, \ldots, x^{(s)}, \infty, \ldots, \infty); \qquad (5.14)$$

$$f_{\xi_1}(x^{(1)}, \ldots, x^{(s)}) = \int\limits_{x^{(s+1)}} \int\limits_{x^{(s+2)}} \ldots \int\limits_{x^{(p)}} f_\xi(x^{(1)}, \ldots$$
$$\ldots, x^{(s)}, x^{(s+1)}, \ldots, x^{(p)})\, dx^{(s+1)} \ldots dx^{(p)}, \qquad (5.15)$$

où le signe $\int\limits_{x^{(q)}}$ indique que l'intégration est étendue à l'ensemble tout entier des valeurs possibles de $\xi^{(q)}$ (comparer avec (5.3) et (5.3′)).

La *réalisation empirique des formules* (5.14) et (5.15) est très simple : nous ne soumettrons à un traitement statistique par les formules (5.12) et (5.13) qu'*une partie seulement* des coordonnées, plus exactement, les points $(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(s)})$, $i = 1, 2, \ldots, n$, et laisserons de côté les valeurs observées des autres coordonnées $x^{(s+1)}, \ldots, x^{(p)}$. Géométriquement, cela signifie que nous projetons les points-observations de l'espace $(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ sur l'espace engendré par les s premières coordonnées (pour $p = 2$, cette opération consiste à projeter n points-observations du plan $x^{(1)}Ox^{(2)}$ sur l'axe $Ox^{(1)}$).

La *densité de probabilité conditionnelle* $f_\xi(x^{(1)}, x^{(2)}, \ldots, x^{(s)} \mid \xi_2 = C)$ du sous-vecteur aléatoire $\xi_1 = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(s)})$, sachant que les valeurs de l'autre sous-vecteur $\xi_2 = (\xi^{(s+1)}, \xi^{(s+2)}, \ldots, \xi^{(p)})$ sont fixées au niveau $C = (c^{(s+1)}, c^{(s+2)}, \ldots, c^{(p)})$ (c'est-à-dire sachant que $\xi^{(s+1)} = c^{(s+1)}, \ldots, \xi^{(p)} = c^{(p)}$), se définit comme les probabilités conditionnelles $p_{i.}$ et $p_{.j}$ (cf. (5.4) et (5.4′)) à l'aide du théorème

---

vailler sur un bien plus grand nombre de données empiriques. En définissant les conditions que doit remplir la taille de l'échantillon, on tiendra compte du fait qu'on aura au total $m_1 m_2 \ldots m_p$ *hyperparallélépipèdes de groupement*, où m_q est le nombre d'*intervalles* de groupement obtenus par quantification de l'intervalle de variation des observations par rapport à la q-ième composante, c'est-à-dire la variable $x^{(q)}$ $(q = 1, 2, \ldots, p)$.

de multiplication des probabilités (cf. n° 4.1.3, (4.11')):

$$f_{\xi_1}(x^{(1)}, \ldots, x^{(s)} \mid \xi^{(s+1)} = c^{(s+1)}, \ldots, \xi^{(p)} = c^{(p)}) =$$

$$= \frac{f_\xi(x^{(1)}, \ldots, x^{(s)}, c^{(s+1)}, \ldots, c^{(p)})}{f_{\xi_2}(c^{(s+1)}, c^{(s+2)}, \ldots, c^{(p)})}. \qquad (5.16)$$

De façon analogue,

$$f_{\xi_2}(x^{(s+1)}, \ldots, x^{(p)} \mid \xi^{(1)} = c^{(1)}, \ldots, \xi^{(s)} = c^{(s)}) =$$

$$= \frac{f_\xi(c^{(1)}, \ldots, c^{(s)}, x^{(s+1)}, \ldots, x^{(p)})}{f_{\xi_1}(c^{(1)}, \ldots, c^{(s)})}. \qquad (5.16')$$

Les dénominateurs des seconds membres de (5.16) et (5.16') sont les densités de probabilité marginales respectivement des sous-vecteurs $\xi_2 = (\xi^{(s+1)}, \ldots, \xi^{(p)})$ et $\xi_1 = (\xi^{(1)}, \ldots, \xi^{(s)})$, calculées à l'aide de (5.15).

Il existe une différence fondamentale entre la densité de probabilité marginale $f_{\xi_1}(x^{(1)}, \ldots, x^{(s)})$ et la densité conditionnelle $f_{\xi_1}(x^{(1)}, \ldots, x^{(s)} \mid \xi^{(s+1)} = c^{(s+1)}, \ldots, \xi^{(p)} = c^{(p)})$, bien que toutes deux décrivent la répartition d'un même ensemble de variables $\xi^{(1)}, \ldots, \xi^{(s)}$: la première densité ne dépend pas des valeurs prises par les autres composantes de la variable aléatoire vectorielle analysée, et son analogue empirique se construit à l'aide de toutes les observations tronquées $\{\widetilde{X}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(s)})\}_{i=1,\ldots,n}$, tandis que la densité conditionnelle dépend essentiellement des niveaux $c^{(s+1)}, \ldots, c^{(p)}$ de fixation des valeurs des autres composantes $\xi^{(s+1)}, \ldots, \xi^{(p)}$, et son analogue empirique se construit à l'aide des observations $X_1, X_2, \ldots, X_n$ dont les p-s dernières coordonnées satisfont au moins approximativement la condition $\{\xi^{(s+1)} = c^{(s+1)}, \ldots, \xi^{(p)} = c^{(p)}\}$.

La figure 5.7 représente le graphique de la densité de probabilité de la loi normale à deux dimensions (qui est décrite au § 6.1). Sur cette figure sont également tracées les sections de la surface représentative de la densité par les plans $x^{(1)} = c$, c'est-à-dire par des plans perpendiculaires à l'axe $Ox^{(1)}$. Les sections sont à un facteur de normalisation près des lois (l'une d'elles est indiquée par une flèche) caractérisant la répartition conditionnelle de la composante $\xi^{(2)}$ sachant que $\xi^{(1)} = c$. La droite OA décrit les variations de la valeur modale de la variable aléatoire $\xi^{(2)}$ dans la répartition conditionnelle de $\xi^{(2)}$ (sachant que $\xi^{(1)} = c$) en fonction de la valeur fixée c.

L'*indépendance statistique de variables aléatoires* $\xi_1, \xi_2, \ldots, \xi_k$ (qui peuvent être discrètes ou continues, scalaires ou vectorielles)

se définit à l'aide de la notion d'indépendance d'un système d'événements (cf. (4.12) et (4.13)). On dit que des *variables aléatoires* $\xi_1, \xi_2, \ldots, \xi_k$ *sont statistiquement indépendantes si pour tous domaines*

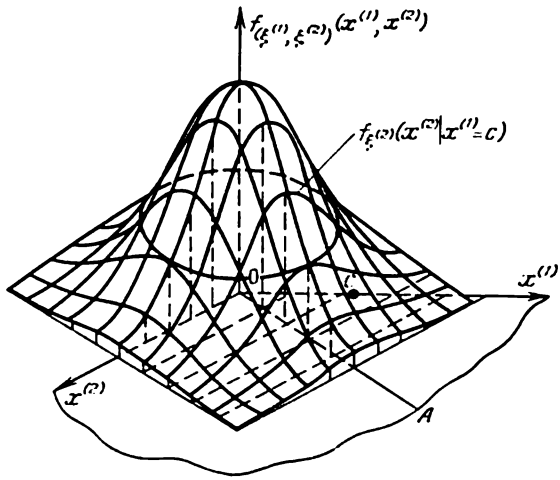


Fig. 5.7. Surface de la densité de probabilité à deux dimensions de la loi normale tronquée

(*mesurables*) $A_1, A_2, \ldots, A_k$ de leurs valeurs possibles, on a les relations

$$P\{\xi_1 \in A_1, \xi_2 \in A_2, \ldots, \xi_k \in A_k\} = \\ = P\{\xi_1 \in A_1\} \ldots P\{\xi_k \in A_k\}. \qquad (5.17)$$

En termes de probabilités $p_{i_1 i_2 \ldots i_k}$ (pour les variables aléatoires discrètes) et de densités de probabilité $f_{\xi_1 \ldots \xi_k}(X^{(1)}, \ldots, X^{(k)})$ (pour les variables continues), la condition (5.17) est équivalente à :

$$p_{i_1 i_2 \ldots i_k} = P\{\xi_1 = X_{i_1}^{(1)\circ}, \ldots, \xi_k = X_{i_k}^{(k)\circ}\} = \\ = P\{\xi_1 = X_{i_1}^{(1)\circ}\} \ldots P\{\xi_k = X_{i_k}^{(k)\circ}\}; \qquad (5.17')$$

$$f_{(\xi_1 \ldots \xi_k)}(X^{(1)}, \ldots, X^{(k)}) = f_{\xi_1}(X^{(1)}) f_{\xi_2}(X^{(2)}) \ldots f_{\xi_k}(X^{(k)}). \qquad (5.17'')$$

5.6. Principales caractéristiques numériques des variables aléatoires et leurs analogues empiriques

Ainsi, la loi de probabilité qui nous intéresse peut être décrite de *façon exhaustive* soit par le polygone de fréquence dans le cas discret, soit par la fonction de répartition ou la densité de probabilité dans le cas continu. Mais en pratique, *on peut se contenter d'une information*

bien plus modeste sous forme de quelques caractéristiques permettant d'estimer des propriétés telles que le centre de groupement des valeurs de la variable aléatoire étudiée, la mesure de leur dispersion aléatoire, le degré d'interdépendance des diverses composantes de la variable aléatoire vectorielle analysée. Ainsi, quand on étudie la loi de répartition du salaire, on s'intéresse en premier lieu au salaire moyen et à l'une des mesures de sa dispersion aléatoire: le *coefficient de différentiation* ou *variance*. De plus, la plupart des lois de répartition utilisées dans les applications statistiques (les lois binomiale, de Poisson, de Pareto, normale, log-normale, exponentielle, etc., cf. chap. 6) peuvent être déterminées sans ambiguïté à l'aide d'une ou deux caractéristiques numériques, par exemple la moyenne et la variance.

5.6.1. Notion d'espérance mathématique et de moment. On considérera diverses fonctions $g(\xi)$ d'une variable aléatoire $\xi = \xi(\omega)$ (si ξ est une variable aléatoire scalaire, il est possible que $g(\xi) \equiv \xi$). Il est évident que la fonction $g(\xi(\omega))$ est une variable aléatoire, puisque c'est une fonction qui est définie sur l'ensemble des événements élémentaires ω. La moyenne de la variable aléatoire $g(\xi)$, prise en tenant compte de la « pondération » correspondant à la loi de probabilité de ξ, s'appelle *espérance mathématique de* $g(\xi)$ et se note $\mathrm{E}g(\xi)$ *). Donc,

si ξ est une variable aléatoire (scalaire ou vectorielle) continue de densité de probabilité (conjointe) $f_\xi(X)$, alors

$$\mathrm{E}g(\xi) = \int g(X) f_\xi(X)\, dX \tag{5.18}$$

(l'intégration est étendue au domaine de toutes les valeurs possibles de la variable ξ);

si ξ est une variable aléatoire discrète prenant les valeurs X_i^0 avec les probabilités p_i ($i = 1, 2, \ldots$), alors

$$\mathrm{E}g(\xi) = \sum_i g(X_i^0) \cdot p_i. \tag{5.18'}$$

L'*analogue empirique* (c'est-à-dire *construit à l'aide des valeurs observées* $X_1, X_2, \ldots, X_n$) de l'espérance mathématique de la fonction $g(\xi)$ est la quantité **)

$$\overline{g(\xi)} = \frac{1}{n} \sum_{i=1}^{n} g(X_i). \tag{5.19}$$

*) Si $g(\xi)$ est une fonction vectorielle, c'est-à-dire que $g(\xi) = (g^{(1)}(\xi), g^{(2)}(\xi), \ldots, g^{(p)}(\xi))$, la *moyenne* pondérée porte sur les composantes.

**) Ici et dans la suite la barre horizontale désignera (sauf indication contraire) la moyenne « arithmétique » ordinaire (sur toutes les valeurs observées) de l'expression qu'elle surmonte.

Tout passage des caractéristiques *théoriques*, c'est-à-dire des caractéristiques calculées d'après la connaissance exacte de la loi de répartition analysée, aux *empiriques* (*d'échantillonnage*) s'explique par le fait que l'échantillon est interprété comme un modèle réduit de la population générale dont les *valeurs possibles* sont les valeurs observées (c'est-à-dire réalisées) $X_1, X_2, \ldots, X_n$ et les probabilités de réalisation de ces dernières, les fréquences respectives de leur apparition dans l'échantillon, c'est-à-dire des quantités égales à $\frac{1}{n}$.

Les fonctions $g_k(\xi) = \xi^k$ et $g_k^0(\xi) = (\xi - \mathrm{E}\xi)^k$ $(k = 1, 2, \ldots)$ jouent un rôle important dans la théorie et la pratique des recherches statistiques. Les espérances mathématiques des fonctions $g_k(\xi)$ et $g_k^0(\xi)$ s'appellent respectivement *moment initial et moment centré d'ordre k* de la variable aléatoire ξ.

Donc, si pour un entier $k > 0$, la fonction x^k est intégrable de poids $f_\xi(x)$ (est sommable de poids p_i) sur le domaine des valeurs possibles de ξ, la quantité

$$m_k = \mathrm{E}\xi^k = \begin{cases} \int x^k f_\xi(x)\,\mathrm{d}x & \text{si } \xi \text{ est continue;} \\ \sum\limits_i (x_i^0)^k \cdot p_i & \text{si } \xi \text{ est discrète} \end{cases} \tag{5.20}$$

s'appelle moment initial d'ordre k ou simplement k-moment de ξ ou de sa loi, et l'on dit que ce moment existe ou est fini.

Il est évident que si le moment m_k existe, il en est de même du moment *centré*

$$m_k^{(0)} = \mathrm{E}(\xi - m_1)^k =$$

$$= \begin{cases} \int (x - m_1)^k f(x)\,\mathrm{d}x & \text{si } \xi \text{ est continue;} \\ \sum\limits_i (x_i^0 - m_1)^k p_i & \text{si } \xi \text{ est discrète.} \end{cases} \tag{5.21}$$

En développant $(x - m_1)^k$ sous le signe d'intégration (ou de la somme), on établit sans peine les relations suivantes entre les moments initiaux et centrés :

$$\begin{aligned} m_1^{(0)} &= 0; \\ m_2^{(0)} &= m_2 - m_1^2; \\ m_3^{(0)} &= m_3 - 3m_1m_2 + 2m_1^3; \\ m_4^{(0)} &= m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 \\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{5.22}$$

(on s'est limité aux quatre premiers moments).

Les *analogues empiriques des moments initiaux et centrés* (les *moments empiriques*) se déduisent sans peine de (5.20) et (5.21) en

tenant compte de (5.19):

$$\hat{m}_k(n) = \frac{1}{n} \sum_{i=1}^{n} x_i^k; \tag{5.20'}$$

$$\hat{m}_k^0(n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{m}_1(n))^k. \tag{5.21'}$$

Enfin, dans l'étude des variables aléatoires vectorielles $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ un rôle important est occupé par les fonctions *vectorielles* $g(\xi) = (g^{(1)}(\xi), \ldots, g^{(p^2)}(\xi))$ dont les composantes sont tous les produits deux à deux des composantes centrées de ξ, c'est-à-dire les éléments de la matrice

$$Q = (q_{ij})_{i,\, j=1,\, \ldots,\, p},$$

où

$$q_{ij} = (\xi^{(i)} - m_1^{(i)})(\xi^{(j)} - m_1^{(j)}).$$

Les espérances mathématiques des éléments q_{ij} s'appellent généralement *moments mixtes d'ordre deux* ou *covariances* de ξ, et la matrice

$$\sum (\mathrm{E}q_{ij})_{i,\, j=1,\, \ldots,\, p}, \tag{5.23}$$

composée des covariances

$$\operatorname{cov}(\xi^{(i)}, \xi^{(j)}) = \mathrm{E}q_{ij} = \mathrm{E}\{(\xi^{(i)} - m_1^{(i)})(\xi^{(j)} - m_1^{(j)})\} = \sigma_{ij}, \tag{5.24}$$

matrices des covariances de ξ.

Par définition, les matrices des covariances sont toutes *symétriques* (c'est-à-dire que $\sigma_{ij} = \sigma_{ji}$); il est immédiat de prouver qu'elles sont semi-définies positives. En effet, en prenant une suite de réels $t_1, t_2, \ldots, t_p$ quelconques et en tenant compte du fait que la quantité positive $\mathrm{E}\left(\sum_{i=1}^{p} t_i (\xi^{(i)} - m_1^{(i)})\right)^2$ peut être représentée par la *forme quadratique* $\sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} \cdot t_i \cdot t_j$, on établit la semi-définition positive de la matrice Σ.

5.6.2. Caractéristiques du centre de groupement des valeurs de la variable aléatoire. Pour caractériser le centre de groupement des valeurs de la variable aléatoire étudiée, on se sert en statistique de quelques types de moyennes, du mode et de la médiane. Décrivons ces caractéristiques numériques.

La *moyenne théorique* $\mathrm{E}\xi$ de la variable aléatoire ξ se définit comme le moment initial d'ordre un ou, ce qui est équivalent, comme l'espé-

rance mathématique (cf. (5.18) et (5.20)):

$$\mathrm{E}\xi = m_1 = \begin{cases} \int x \cdot f_\xi(x)\,dx & \text{si } \xi \text{ est continue;} \\ \sum_i x_i^0 \cdot p_i & \text{si } \xi \text{ est discrète.} \end{cases} \tag{5.25}$$

La moyenne $\mathrm{E}\xi$ est indiscutablement la plus importante et la plus couramment utilisée des caractéristiques du centre de groupement des valeurs de la variable aléatoire. En pratique, c'est-à-dire quand on la calcule approximativement à l'aide des valeurs empiriques $x_1, x_2, \ldots, x_n$, on la remplace par la *moyenne empirique* (d'échantillonnage) (cf. (5.20')):

$$\bar{x}(n) = \hat{m}_1(n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Les principales propriétés de l'espérance mathématique résultent directement de sa définition:

a) $\mathrm{E}c = c$, où c est une variable non aléatoire quelconque;

b) $\mathrm{E}(c\xi) = c \cdot \mathrm{E}\xi$;

c) $\mathrm{E}(\xi_1 + \xi_2 + \ldots + \xi_N) = \mathrm{E}\xi_1 + \mathrm{E}\xi_2 + \ldots + \mathrm{E}\xi_N$;

d) $\mathrm{E}(\xi \cdot \eta) = \mathrm{E}\xi \cdot \mathrm{E}\eta$ si les variables aléatoires ξ et η sont indépendantes (cf. n° 5.5.3).

La *moyenne géométrique* (*théorique*) $G(\xi)$ d'une variable aléatoire ξ se définit (pour des valeurs possibles de $\xi > 0$) à l'aide de la formule

$$G(\xi) = e^{\mathrm{E}(\ln \xi)},$$

où $e \approx 2{,}71828\ldots$ L'analogue empirique de $G(\xi)$, la moyenne géométrique $\hat{G}_n(\xi)$, se calcule sur les valeurs d'échantillonnage $x_1, x_2, \ldots$ $\ldots, x_n$ à l'aide de la formule

$$\hat{G}_n(\xi) = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

On démontre que la moyenne géométrique $G(\xi)$ et son analogue empirique $\hat{G}_n(\xi)$ sont toujours strictement inférieurs respectivement à $\mathrm{E}\xi$ et $\bar{x}(n)$.

La moyenne géométrique s'applique au calcul de la vitesse de variation des variables et, notamment, dans les cas où la variable est directement proportionnelle à la valeur qu'elle prend à l'instant considéré (par exemple, effectif d'une population) ou dans les cas où l'on a affaire à une moyenne de rapports, par exemple, dans le calcul des indices des prix.

La *moyenne harmonique* (*théorique*) $H(\xi)$ d'une variable aléatoire ξ est définie (uniquement pour les variables ξ prenant des valeurs

strictement positives) par la relation

$$H(\xi) = \frac{1}{\mathrm{E}\left(\frac{1}{\xi}\right)}.$$

L'*analogue empirique* $\hat{H}_n(\xi)$ se calcule sur les données d'échantillonnage $x_1, x_2, \ldots, x_n$ à l'aide de la formule

$$\hat{H}_n(\xi) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}.$$

La moyenne harmonique est souvent inférieure à la moyenne géométrique et *a fortiori* à la moyenne arithmétique. Le domaine d'application de la moyenne harmonique est fort limité. On s'en sert en particulier en économie pour analyser les normes moyennes de temps, ainsi que dans le calcul de certains indices.

La *valeur modale* (ou simplement *mode*) x_{mod} d'une variable aléatoire se définit comme une valeur possible qui maximise la densité de probabilité $f_\xi(x)$ (dans le cas continu) ou la probabilité $P\{\xi = x\}$ (dans le cas discret). Donc, le mode est en quelque sorte la valeur la plus typique de la variable aléatoire, celle qui se réalise le plus souvent (dans les expériences ou les observations), bref, la valeur qui est réellement à la « mode » *). Le calcul pratique d'une valeur approchée du mode sur les données d'échantillonnage implique la construction et l'analyse des histogrammes et polygones de fréquence correspondants (cf. §§ 5.5, 10.3).

La *médiane* $x_{\text{méd}}$ de la variable aléatoire étudiée se définit comme la valeur moyenne probable, c'est-à-dire une valeur qui jouit de la propriété suivante: la probabilité que la variable aléatoire ξ prenne des valeurs $> x_{\text{méd}}$ est égale à la probabilité qu'elle prenne des valeurs $< x_{\text{méd}}$. Pour les variables dont la densité de probabilité est continue, il est évident que

$$P\{\xi > x_{\text{méd}}\} = P\{\xi < x_{\text{méd}}\} = 0{,}5,$$

et la médiane peut être définie comme une valeur $x_{\text{méd}}$ de l'axe des valeurs possibles (axe des abscisses) telles que la droite parallèle à l'axe des ordonnées passant par $x_{\text{méd}}$ partage la surface située sous la courbe de densité en deux parties égales (fig. 5.8). Certaines variables aléatoires discrètes ne vérifient pas exactement la condition

*) Le mode est une caractéristique naturelle du centre de groupement des valeurs de la variable aléatoire uniquement dans les cas de répartitions unimodales (à un sommet). Les répartitions multimodales (à plusieurs sommets) traduisent l'extrême inhomogénéité de la population analysée. Leur étude présente un intérêt immédiat pour les *problèmes de classification* des objets et des observations.

formulée. Dans ce cas, la médiane se définit comme un nombre $x_{\text{méd}}$ compris entre deux valeurs possibles voisines $x^0_{i(0,5)}$ et $x^0_{i(0,5)+1}$ telles que $F_\xi(x^0_{i(0,5)}) < 0{,}5$, mais $F_\xi(x^0_{i(0,5)+1}) \geqslant 0{,}5$.

Pour déterminer une valeur (*d'échantillonnage*) approchée de la médiane $\hat{x}_{\text{méd}}(n)$, on range les valeurs observées $x_1, x_2, \ldots, x_n$ dans l'ordre de grandeur croissante (ou en *série variationnelle*, cf. n° 5.6.4), puis pour $\hat{x}_{\text{méd}}(n)$ on prend le $\frac{1}{2}(n+1)$-ième terme si n est impair, et toute valeur comprise entre le $\frac{1}{2}n$-ième et le $\left(\frac{1}{2}n+1\right)$-ième terme si n est pair.

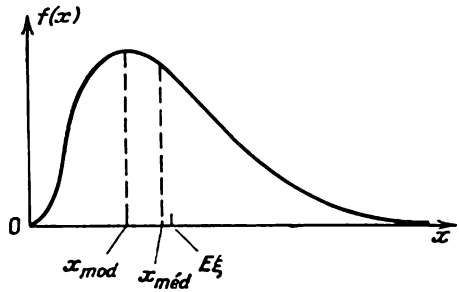


Fig. 5.8. Disposition du mode x_{mod}, de la médiane $x_{\text{méd}}$ et de la moyenne $E\xi$ d'une loi dont la densité $f(x)$ est dissymétrique

Si la densité (ou le polygone de fréquence) est symétrique, la moyenne $E\xi$, le mode x_{mod} et la médiane $x_{\text{méd}}$ sont confondus La situation est différente pour les répartitions non symétriques (cf. fig. 5.8).

5.6.3. Caractéristiques du degré de dispersion d'une variable aléatoire. Chacune des caractéristiques du degré de dispersion décrites

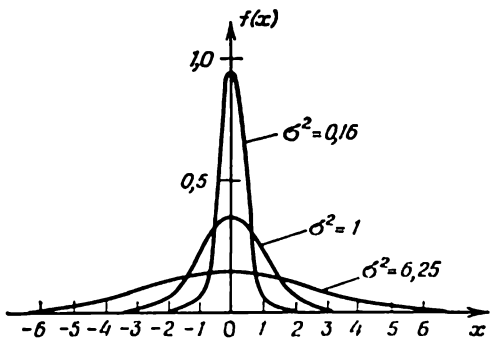


Fig. 5.9. Représentation graphique des densités (de lois normales) de même moyenne (nulle) et de variance différente

plus bas — la variance, l'écart-type et le coefficient de variation — donne une idée sur l'ampleur de l'écart des valeurs de la variable aléatoire par rapport à son centre de groupement. S'agissant de la forme de la courbe de densité, ces caractéristiques décrivent son degré d'« aplatissement » sur l'intervalle de variation de ξ: plus la

valeur de chacune de ces caractéristiques est élevée, plus la courbe de la répartition correspondante est «aplatie» (fig. 5.9).

La variance $V\xi$ d'une variable aléatoire ξ se définit comme le moment centré d'ordre deux de ξ, c'est-à-dire que

$$V\xi = m_2^{(0)} = \begin{cases} \int\limits_X (x - E\xi)^2 f(x)\,dx & \text{si } \xi \text{ est continue;} \\ \sum\limits_i (x_i^{(0)} - E\xi)^2 \cdot p_i & \text{si } \xi \text{ est discrète.} \end{cases} \tag{5.26}$$

La *variance empirique* $s^2(n)$ peut être considérée comme une valeur approchée de la variance théorique:

$$s^2(n) = \hat{m}_2^{(0)}(n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

De la définition de la variance (et des propriétés de l'espérance mathématique) on déduit les propriétés suivantes:

a) $Vc = 0$ (c est une variable non aléatoire);

b) $V(c\xi) = c^2 \cdot V\xi$;

c) $V(a + b\xi) = b^2 \cdot V\xi$ (a et b sont des variables non aléatoires);

d) $V(\xi + \eta) = V\xi + V\eta$ si ξ et η sont indépendantes.

La variance est souvent désignée par σ^2.

L'*écart-type* (ou *écart quadratique moyen*) σ_ξ est défini comme la racine carrée de la variance: $\sigma_\xi = \sqrt{V\xi}$. Il sert au même titre que la variance à caractériser le degré d'éparpillement de la variable aléatoire et s'avère parfois plus commode à l'usage en raison de son homogénéité (relativement aux unités de mesure) avec les diverses caractéristiques du centre de groupement.

La *valeur empirique* de l'écart-type est donnée par la formule *)

$$s(n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Le *coefficient de variation* v_ξ est utilisé dans les cas où le degré de dispersion se décrit naturellement par une relation entre une caractéristique relative et l'espérance mathématique. En particulier,

$$v_\xi = \frac{\sqrt{V\xi}}{E\xi} \cdot 100\,\% = \frac{\sigma_\xi}{E\xi} \cdot 100\,\%,$$

autrement dit, le coefficient de variation est le rapport (en pourcentage) de l'écart-type à l'espérance mathématique. On voit sur la

*) Les expressions « corrigées » $\bar{s}^2 = \frac{n}{n-1} s^2(n)$ et $\bar{s} = \sqrt{\frac{n}{n-1} \cdot s^2(n)}$ (cf. § 8.6) approchent mieux les valeurs théoriques de la variance et de l'écart-type surtout lorsque n n'est pas grand.

définition que v_ξ est une quantité sans dimension. La caractéristique empirique correspondante est définie par la formule

$$\hat{v}_\xi(n) = \frac{s(n)}{\bar{x}(n)} \cdot 100\ \%.$$

5.6.4. Série variationnelle et statistiques de rang. Nous avons signalé plus haut que l'échantillon, c'est-à-dire l'*ensemble* des valeurs observées $x_1, x_2, \ldots, x_n$ de la variable aléatoire ξ, constitue l'information de base à partir de laquelle l'analyste tire ses conclusions sur les propriétés de la population générale et, en particulier, se fait une idée de la fonction de répartition et du polygone de fréquence ou de la densité de la loi de probabilité analysée (cf. §§ 5.5, 10.3 et 10.4). Chaque terme de l'échantillon peut isolément fournir une importante information sur le caractère de la loi de probabilité si les observations sont rangées dans leur ordre de grandeur croissante. Ainsi, les valeurs $x_{\min}(n)$ et $x_{\max}(n)$ donnent une idée approximative sur les marges de variation des valeurs possibles de la variable aléatoire ξ, la différence $x_{\max}(n) - x_{\min}(n)$, sur le degré d'éparpillement des valeurs observées; le terme médian de la série variationnelle des observations — la médiane $\hat{x}_{\text{méd}}(n)$ — caractérise le centre de groupement des observations de la variable aléatoire, etc. Tout ceci nous suggère un traitement spécial pour les observations classées par ordre de grandeur croissante.

Soient $x_1, x_2, \ldots, x_n$ un échantillon composé de n observations *indépendantes* *) d'une variable aléatoire ξ de fonction de répartition

*) Expliquons ce qu'on entend par observations *indépendantes* à l'aide de la notion d'indépendance statistique des variables aléatoires. Dans cet ouvrage et dans d'autres, l'échantillon et les observations admettent selon le contexte deux interprétations différentes qui généralement ne sont pas spécifiées et, pour simplifier l'exposé, ne sont pas distinguées par des notations différentes. Jusque-là nous nous sommes contentés de la première interprétation dans le cadre de laquelle par $x_1, x_2, \ldots, x_n$ on comprend les valeurs de ξ *réellement observées* dans l'expérience considérée, bref, des nombres concrets. Dorénavant nous appliquerons parfois aux séries initiale et ordonnée d'observations la deuxième interprétation (l'interprétation *hypothétique*) dans laquelle $x_1, x_2, \ldots, x_n$ sont comprises comme les *valeurs que nous aurions pu obtenir en n répétitions* de l'expérience (l'observation) sous le système de conditions induisant la population générale de la variable ξ. Il est évident que dans cette interprétation, la suite de symboles $(x_1, x_2, \ldots, x_n)$ doit être traitée comme une *variable aléatoire à n dimensions* dont les composantes x_i admettent toutes la même loi de probabilité marginale $f_\xi(x)$. D'après la convention adoptée, pour désigner ces données empiriques nous aurions dû utiliser d'autres notations, par exemple $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)})$, où $f_{\xi^{(i)}}(x) = f_\xi(x)$. Ce procédé de notations qui est plus soigneux (mais qui ne sera pas utilisé dans la suite) permet de définir rigoureusement la notion d'indépendance des observations: les observations hypothétiques $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)}$ de la variable aléatoire ξ sont indépendantes si leur densité de probabilité conjointe $f_{(\xi^{(1)}, \ldots, \xi^{(n)})}(x_1, \ldots, x_n)$ peut être mise sous la forme $f_{(\xi^{(1)}, \ldots, \xi^{(n)})}(x_1, \ldots, x_n) = f_\xi(x_1) \cdot f_\xi(x_2) \cdot \ldots \cdot f_\xi(x_n)$.

$F_\xi(x)$ et de densité de probabilité $f_\xi(x)$ continues (on se limitera comme toujours uniquement à l'étude de telles variables aléatoires continues).

Si les x_i sont rangées par ordre de grandeur croissante et les termes de la suite strictement croissante ainsi obtenue désignés par $x_{(j)}$, c'est-à-dire que

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)},$$

alors chaque $x_{(j)}$ s'appelle *statistique de rang* et la suite

$$x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)} \tag{5.27}$$

série variationnelle de la variable aléatoire ξ.

L'appareil des statistiques de rang est largement utilisé aussi bien en théorie et en pratique des estimations statistiques de paramètres inconnus et de tests statistiques (notamment, dans la construction d'estimations et de tests stables « distribution free », cf. n° 8.6.4, § 10.3, ainsi que les §§ 11.1, 11.2 et 11.3) que dans la modélisation des systèmes et processus réels (cf., par exemple, [3], [4]). Mais quand on étudie la qualité des estimations, tests et modèles obtenus à l'aide des statistiques de rang, il faut avoir une idée sur leur comportement lors d'une répétition de l'échantillonnage, c'est-à-dire qu'il faut savoir décrire les lois de probabilité dans le cadre de l'interprétation *hypothétique*, interprétation dans laquelle, rappelons-le, les termes de la série variationnelle $x_{(j)}$ sont traités non pas comme des nombres concrets mais comme des variables aléatoires. Et bien que les *termes de la série variationnelle* (5.27) *contrairement à ceux de l'échantillon initial ne soient plus indépendants* (car ordonnés) *et, donc, leurs répartitions marginales ne soient plus identiques et décrites notamment par la même densité de probabilité* $f_\xi(x)$, ils peuvent néanmoins être décrits facilement en termes de densité et de fonction de répartition $F_\xi(x)$.

On établit sans peine que la densité de répartition $f_{x_{(j)}}(x)$ de la statistique de rang $x_{(j)}$ de la série variationnelle (5.27) a pour expression

$$f_{x_{(j)}}(x) = n \cdot C_{n-1}^{j-1} F_\xi^{j-1}(x)\,[1 - F_\xi(x)]^{n-j} f_\xi(x) \tag{5.28}$$

(pour prouver cette formule, on se sert de la loi de probabilité polynomiale, cf. § 6.1).

Le comportement des termes de la série variationnelle est décrit de façon exhaustive par la densité de probabilité conjointe $f_{(x_{(1)},\, x_{(2)}, \ldots, x_{(n)})}(x, y, \ldots, t)$. Le calcul d'une telle densité (en fonction de $f_\xi(x)$ et de $F_\xi(x)$) n'est pas compliqué, mais est assez laborieux. Citons quelques exemples de répartitions de statistiques de rang et de fonctions de statistiques de rang les plus fréquemment utilisées en pratique.

La *répartition conjointe des statistiques de rang* $x_{(j)}$ *et* $x_{(q)}$ $(j < q)$ se définit par la densité

$$f_{(x_{(j)}x_{(q)})}(x, y) = K(j, q, n) \cdot F_{\xi}^{j-1}(x)\,[F_{\xi}(y) - \\ - F_{\xi}(x)]^{q-j-1}\,[1 - F_{\xi}(y)]^{n-q} f_{\xi}(x)\, f_{\xi}(y), \quad (5.29)$$

où le coefficient $K(j, q, n)$ est donné par la formule

$$K(j, q, n) = \frac{n!}{(j-1)!\,(q-j-1)!\,(n-q)!}. \quad (5.29')$$

En particulier, pour $j = 1$ et $q = n$, c'est-à-dire pour les termes *extrêmes* de la série variationnelle, on obtient la densité de probabilité

$$f_{(x_{(1)}, x_{(n)})}(x, y) = n(n-1)\,[F_{\xi}(y) - F_{\xi}(x)]^{n-2} f_{\xi}(x)\, f_{\xi}(y). \quad (5.29'')$$

Une importante fonction des statistiques de rang, présente dans de nombreuses applications, est l'*étendue* $\Delta_n = x_{(n)} - x_{(1)}$ qui est utilisée au même titre que la dispersion, l'écart-type et le coefficient de variation (cf. n° 5.6.3) pour caractériser le degré d'éparpillement de la variable aléatoire étudiée. La formule (5.29″) permet d'obtenir la fonction de répartition de l'étendue

$$F_{\Delta_n}(z) = n \int [F_{\xi}(x+z) - F_{\xi}(x)]^{n-1} f_{\xi}(x)\, dx, \quad (5.30)$$

l'intégration étant étendue au domaine des valeurs possibles de la variable aléatoire ξ.

Voyons maintenant un exemple d'application de l'appareil des statistiques de rang à la modélisation des phénomènes économiques. Dans le travail [4] cet appareil est utilisé pour la construction de modèles prévisionnels de répartition des familles et de leurs membres en fonction du revenu par personne, et en particulier pour l'étude de la structure et du caractère des relations entre les répartitions: en fonction du salaire de *tous* les membres actifs ($f_{\xi}(x)$), en fonction uniquement des premiers, des seconds, etc. membres actifs dans les familles de n membres actifs ($f_{x_{(j)}}(x)$, $j = 1, 2, \ldots, n$; $x_{(j)}$ est le j-ième salaire par son montant dans la famille); des familles en fonction du revenu moyen par personne η ($f_{\eta}(x)$). Les modèles décrits dans [4] ont permis en particulier de déterminer avec beaucoup d'exactitude les répartitions $f_{x_{(j)}}(x)$ et $f_{\eta}(x)$ à l'aide de la répartition connue $f_{\xi}(x)$ de tous les membres actifs en fonction du montant de leur salaire et aussi de calculer les diverses caractéristiques de la relation statistique entre, d'une part, le salaire du j-ième membre actif dans la famille $x_{(n-j+1)}$ et le revenu moyen par personne η, et, d'autre part, le salaire d'un membre actif ξ pris au hasard, *sans procéder à des sondages statistiques spéciaux dans les populations géné-*

rales $x_{(n-j+1)}$ (des j-ièmes membres actifs) et η (familles). Ainsi, dans le cas de populations de familles ayant deux membres actifs, les densités de la répartition du salaire *uniquement* des premiers et *uniquement* des seconds membres actifs peuvent être déterminées en fonction de la loi de probabilité du salaire de tous les membres actifs à l'aide des formules:

$$f_{x_{(2)}}(x) = 2F_\xi(x) f_\xi(x);$$

$$f_{x_{(1)}}(x) = 2[1 - F_\xi(x)] f_\xi(x).$$

5.6.5. Quantiles et points de pourcentage d'une répartition. Les méthodes de statistique mathématique, notamment les tests statistiques (cf. § 9.1) et les estimations par intervalles des paramètres inconnus (cf. chap. 8) utilisent largement les notions de quantile $u_q(F)$ d'ordre q et les points de pourcentage $w_Q(F)$ de la répartition $F(x)$.

On appelle *quantile d'ordre* q (ou q-quantile) d'une variable aléatoire continue ξ de fonction de répartition continue $F(x)$ une valeur possible $u_q(F)$ de ξ telle que la probabilité de l'événement $\{\xi < u_q(F)\}$ est égale à q, soit

$$F(u_q) = P\{\xi < u_q\} = q. \tag{5.31}$$

Il est évident que u_q croît avec q $(0 < q < 1)$. Le quantile d'ordre 0,5 (la médiane) caractérise le centre de groupement.

Nous avons vu que la fonction de répartition $F_\xi(x)$ de toute variable aléatoire discrète ξ croît avec x *par sauts*, donc on peut trouver des q pour lesquels il n'existe pas de valeurs possibles u_q satisfaisant exactement l'équation (5.31). Donc, dans le cas discret, un quantile d'ordre q se définit comme un nombre $u_q(F)$ compris entre deux valeurs possibles voisines $x^0_{i(q)}$ et $x^0_{i(q)+1}$ et tel que $F(x^0_{i(q)}) < q$, mais $F(x^0_{i(q)+1}) \geqslant q$.

Les *analogues empiriques* des quantiles théoriques seront de toute évidence les termes de la série variationnelle (les statistiques de rang). De leurs définitions il s'ensuit en particulier que la statistique de rang $x_{(j)}$ est en même temps le quantile empirique d'ordre $(j-1)/n$, puisque la fréquence relative (l'analogue empirique de la probabilité!) des observations $< x_{(j)}$ est justement égale à $(j-1)/n$.

Les points de pourcentage sont souvent utilisés à la place des quantiles auxquels ils sont étroitement liés. On appelle *point de pourcentage* $100\,Q$ $(0 < Q < 100)$ d'une variable aléatoire ξ sa valeur possible w_Q pour laquelle la probabilité de l'événement $\{\xi \geqslant w_Q\}$ est égale à $Q/100$, c'est-à-dire que

$$1 - F(w_Q) = P\{\xi \geqslant w_Q\} = \frac{Q}{100}.$$

Le point de pourcentage se définit *mutatis mutandis* pour le cas discret.

De la définition du quantile et du point de pourcentage on déduit la relation élémentaire:

$$u_q = w_{100(1-q)}. \tag{5.32}$$

Des tables spéciales de quantiles et de points de pourcentage ont été composées pour les lois de probabilité les plus fréquemment utilisées en statistique (cf. chap. 6). Il est évident qu'il suffit de connaître une de ces tables, car si l'on demande, par exemple, de trouver le quantile d'ordre 0,9 de la loi normale, il suffit, en vertu de (5.32), de déterminer le point de pourcentage 10 de cette même loi.

La figure 5.10 illustre de façon suggestive la signification géométrique des notions introduites. Ici $q = \int_{-\infty}^{u_q} f_\xi(x)\,dx$; $\frac{Q}{100} = \int_{w_Q}^{\infty} f_\xi(x)\,dx$; $y = f_\xi(x)$.

Les quantiles qui sont essentiellement des instruments statistiques auxiliaires peuvent parfois jouer le rôle de principales caractéristiques ou de paramètres de la loi de probabilité étudiée. Ainsi, les *coefficients quantiles* (*d'ordre q*) *de différenciation* $K_d(q)$ qui sont définis par la relation

$$K_d(q) = \frac{u_1 - q}{u_q}$$
$$(0 < q \leqslant 0{,}25)$$

(les plus usités sont les *coefficients déciles* de différenciation qui correspondent à $q = 0{,}1$) sont largement utilisés pour caractériser le degré de dispersion des salaires et des revenus. Les quantiles et les points de pourcentage servent également à représenter les bornes *pratiques* de l'intervalle de variation de la variable étudiée: ainsi, par exemple, les quantiles d'ordre 0,005 et 0,995 définissent parfois respectivement les niveaux minimal et maximal du salaire dans l'échelle respective des indices.

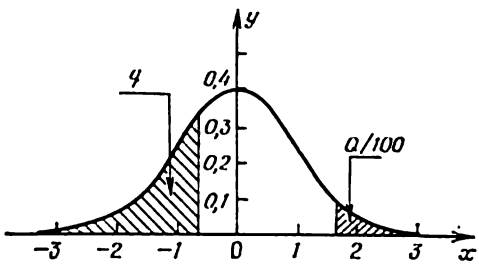


Fig. 5.10. Interprétation géométrique du quantile u_q et du point w_Q de pourcentage $100Q$: cas de la loi normale réduite, $q = 0{,}25$ (resp. $u_{0,25} = -0{,}675$) et $Q = 5\,\%$ (resp. $w_{5\%} = 1{,}65$)

5.6.6. Dissymétrie et aplatissement. On voit sur la formule (5.21), qui définit les moments centrés, que si la densité $f_\xi(x)$ (ou la suite de probabilités $P\{\xi = x_i^0\}$) est symétrique par rapport à la valeur moyenne $m_1 = \mathrm{E}\xi$ (c'est-à-dire que $f(m_1 - x) \equiv f(m_1 + x)$), alors

tous les moments centrés impairs $m_{2k+1}^{(0)}$ (s'ils existent) sont nuls. Donc, tout moment centré impair non nul peut être traité comme la caractéristique de la *dissymétrie* de la répartition correspondante. La plus élémentaire de ces caractéristiques, $m_3^{(0)}$, a été prise pour élément de base dans le calcul du *coefficient de dissymétrie* β_1

$$\beta_1 = \frac{m_3^{(0)}}{(m_2^{(0)})^{3/2}} = \frac{E(\xi - m_1)^3}{[E(\xi - m_1)^2]^{3/2}} \tag{5.33}$$

qui caractérise le degré d'aplatissement de la répartition. La normalisation moyennant la division de $m_3^{(0)}$ par $(m_2^{(0)})^{3/2}$ a été introduite de manière à ce que cette caractéristique ne dépende pas du choix des unités de mesure de la variable aléatoire étudiée: la formule (5.33) définit une caractéristique *sans dimension* du degré d'aplatissement de la répartition, c'est-à-dire invariante par rapport aux unités de mesure de ξ.

Donc, toutes les répartitions symétriques auront un coefficient de dissymétrie nul (cf. fig. 5.5, 5.6, 5.9, 5.10), tandis que les répartitions dont la courbe de densité de probabilité s'étire vers la droite (resp. la gauche), un coefficient strictement *positif* (cf. fig. 5.8) (resp. strictement *négatif*). Le coefficient de dissymétrie empirique $\hat{\beta}_1(n)$ se calcule en fonction des moments centrés empiriques d'ordre deux et trois à l'aide de la formule

$$\hat{\beta}_1(n) = \frac{\hat{m}_3^{(0)}(n)}{(\hat{m}_2^{(0)}(n))^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}. \tag{5.33'}$$

Le comportement de la densité (du polygone) de la répartition au voisinage de sa valeur modale définit la forme géométrique de la courbe correspondante au voisinage de son maximum. Le coefficient d'aplatissement β_2 est une caractéristique utile dans la résolution de nombreux problèmes, par exemple, dans la détermination de la forme générale de la répartition étudiée ou dans son approximation par certains développements spéciaux (voir, par exemple, la représentation des répartitions par des séries de Gram-Charlier et d'Edgeworth [40], p. 246-256). Cette caractéristique se définit à l'aide de la relation:

$$\beta_2 = \frac{m_4^{(0)}}{(m_2^{(0)})^2} - 3 = \frac{E(\xi - m_1)^4}{[E(\xi - m_1)^2]^2} - 3. \tag{5.34}$$

Nous verrons plus bas que la répartition normale de Gauss pour laquelle $\beta_2 = 0$ est une sorte de référence pour mesurer le degré d'aplatissement d'une courbe. En principe, les répartitions dont les courbes de densité (les polygones) sont plus pointues ont un coeffi-

cient d'aplatissement strictement positif, les moins pointues, un coefficient d'aplatissement strictement négatif (fig. 5.11) *).

Le *coefficient d'aplatissement empirique* se calcule à l'aide de la formule

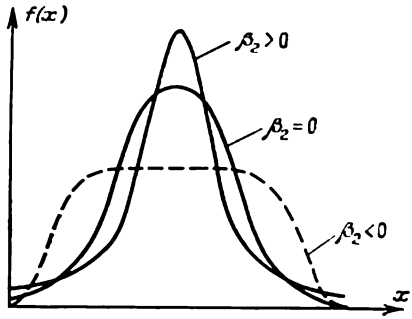


Fig. 5.11. Exemples de densité pour un coefficient d'aplatissement >0, <0 et =0

$$\hat{\beta}_2(n) = \frac{\hat{m}_4^{(0)}(n)}{(\hat{m}_2^{(0)}(n))^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} - 3. \tag{5.34'}$$

5.6.7. Principales caractéristiques des répartitions multidimensionnelles (covariance, corrélation, variance généralisée, etc.). Si en dimension un, et dans une certaine mesure, en dimension deux, l'analyste a encore la possibilité d'utiliser des lois de probabilité modèles convenables (cf. chap. 6), en dimension $p > 2$, il doit se contenter de la seule information que lui fournissent *les deux premiers moments* **): le *vecteur des moyennes*

$$M_1 = \begin{pmatrix} m_1^{(1)} \\ m_1^{(2)} \\ \vdots \\ m_1^{(p)} \end{pmatrix} = \begin{pmatrix} \mathrm{E}\xi^{(1)} \\ \mathrm{E}\xi^{(2)} \\ \vdots \\ \mathrm{E}\xi^{(p)} \end{pmatrix} \tag{5.35}$$

et la *matrice des covariances*

$$\Sigma = \begin{pmatrix} \sigma_{11}\sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21}\sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots \\ \sigma_{p1}\sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}, \tag{5.36}$$

*) En vérité on peut exhiber des exemples violant cette loi (cf. [76]).

**) Diverses tentatives d'introduction et d'utilisation pratique des analogues multidimensionnels des coefficients de dissymétrie et d'aplatissement ont été entreprises. Ainsi Mardia (cf. Mardia K.V. *Measures of multivariate skewness and kurtosis with application.* — Biometrika, vol. 57, 1970) utilise ces analogues pour étudier l'influence des écarts de la répartition analysée par rapport à la loi normale sur les propriétés de la statistique T^2 de Hotelling (cf. § 11.2). Yu. Tiourine a proposé d'autres variantes d'analogues multidimensionnels des coefficients de dissymétrie et d'aplatissement pour le test statistique de la normalité multidimensionnelle de la population générale analysée (cf. *Théorie des probabilités et ses applications.* 1973, t. XVIII, n° 3). Mais les moments d'ordre >2 ne se sont pas avérés être un instrument d'étude efficace et encore moins nécessaire, aussi sont-ils peu sollicités pour le traitement statistique des données multidimensionnelles.

où les moyennes $m_1^{(i)}$ des composantes $\xi^{(i)}$ sont définies à l'aide de la formule (5.25) au moyen des densités marginales (resp. des polygones de fréquence) des variables aléatoires $\xi^{(i)}$, et les covariances $\sigma_{jk} = \mathbf{E}\{(\xi^{(j)} - m_1^{(j)})(\xi^{(k)} - m_1^{(k)})\}$ se calculent au moyen des densités marginales (resp. des polygones de fréquence) du couple de variables aléatoires $(\xi^{(j)}, \xi^{(k)})$.

Les analogues empiriques du vecteur M_1 des moyennes et de la matrice des covariances Σ, c'est-à-dire les caractéristiques calculées directement avec les données d'échantillonnage $X_1, X_2, \ldots, X_n$ sont respectivement le *vecteur des moyennes empiriques*

$$\overline{X}(n) = \hat{M}_1(n) = \begin{pmatrix} \overline{x}^{(1)}(n) \\ \overline{x}^{(2)}(n) \\ \vdots \\ \overline{x}^{(p)}(n) \end{pmatrix} \tag{5.35'}$$

et la *matrice des covariances empiriques*

$$\mathbf{S}(n) = \hat{\Sigma}(n) = \begin{pmatrix} s_{11}(n) & s_{12}(n) & \ldots & s_{1p}(n) \\ s_{21}(n) & s_{22}(n) & \ldots & s_{2p}(n) \\ \ldots & \ldots & \ldots & \ldots \\ s_{p1}(n) & s_{p2}(n) & \ldots & s_{pp}(n) \end{pmatrix}, \tag{5.36'}$$

où $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ est une observation multidimensionnelle, les composantes des moyennes empiriques $\overline{x}^{(k)}(n)$ se calculent à l'aide de la formule (5.35'), et les éléments $s_{jk}(n)$ de la matrice des covariances $\mathbf{S}(n)$, à l'aide de la formule

$$s_{jk}(n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^{(j)} - \overline{x}^{(j)})(x_i^{(k)} - \overline{x}^{(k)}).$$

Comme en dimension un, le vecteur des moyennes est la principale caractéristique du centre de groupement des observations de la variable aléatoire multidimensionnelle ξ (dans l'espace à p dimensions de ses valeurs possibles).

La matrice des covariances Σ caractérise les propriétés suivantes de la variable aléatoire multidimensionnelle ξ.

1. Le *degré de dispersion aléatoire* par rapport à chaque composante et par rapport à la variable ξ tout entière. Il est immédiat de voir que les éléments diagonaux σ_{ii} de la matrice Σ définissent les variances marginales des composantes $\xi^{(i)}$, c'est-à-dire le degré de dispersion aléatoire des valeurs de la variable aléatoire à une dimension $\xi^{(i)}$. Ainsi

$$\sigma_{ii} = \mathbf{E}(\xi^{(i)} - m^{(i)})^2 = \mathbf{V}\xi^{(i)}.$$

L'analogue multidimensionnel de la variance est la valeur du déterminant de la matrice des covariances. On l'appelle *variance généralisée* de la variable aléatoire multidimensionnelle ξ:

$$V_g\xi = \det(\Sigma). \tag{5.37}$$

Une autre caractéristique du degré de dispersion des valeurs de la variable aléatoire multidimensionnelle ξ, souvent utilisée dans les calculs, est la trace de la matrice des covariances Σ, c'est-à-dire la somme de ses éléments diagonaux

$$\mathrm{Tr}(\Sigma) = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}. \tag{5.38}$$

De la semi-définition positive de la matrice Σ (cf. n° 5.6.1) et de la signification des éléments diagonaux σ_{ii}, il s'ensuit que les quantités définies par les formules (5.37) et (5.38) seront toujours positives.

Les analogues empiriques de la variance généralisée (5.37) et de la trace de la matrice Σ (5.38) sont respectivement la *variance généralisée empirique*

$$\hat{V}_g\xi = \det S(n) \tag{5.37'}$$

et la *trace de la matrice des covariances empirique*

$$\mathrm{Tr}(S(n)) = s_{11}(n) + s_{22}(n) + \ldots + s_{pp}(n). \tag{5.38'}$$

Eclaircissons la *signification géométrique* de la variance généralisée (cf., par exemple, [16]). Au sujet de la variance généralisée théorique, on peut dire que si, par exemple, la variable aléatoire multidimensionnelle étudiée suit une loi normale (cf. chap. 6) et si P_0 est une probabilité donnée, alors le volume du domaine (entourant le centre de groupement M_1), dans lequel les valeurs de ξ tombent avec la probabilité P_0, est proportionnel à $\sqrt{V_g\xi}$ (ce volume est proportionnel aussi à un facteur dépendant de la dimension p et proportionnel, en outre, à un nombre défini en fonction de la probabilité donnée P_0). On peut donner également une interprétation géométrique à la variance généralisée empirique dans l'espace à p dimensions des observations $X_1, \ldots, X_n$. A cet effet, considérons dans cet espace tous les parallélépipèdes construits de la manière suivante. Pour supports des arêtes de chaque parallélépipède prenons p à p les vecteurs d'origines p points parmi $X_1, \ldots, X_n$ et d'extrémité le point $\bar{X}$ Il s'avère que la somme des carrés des volumes de tous ces parallélépipèdes sera proportionnelle à la valeur de la variance généralisée empirique $\hat{V}_g\xi$ (le coefficient de proportionnalité étant égal à $1/(n-1)^p$).

2. Le *caractère* et la *structure des relations statistiques* liant les composantes de la variable multidimensionnelle analysée peuvent également être décrits en termes de matrice des covariances. Mais, dans ce cas, il est plus commode de passer à une matrice des covarian-

ces *normalisée* d'une certaine manière: la matrice *des corrélations*

$$\mathbf{R}=\begin{pmatrix} r_{11}r_{12} & \dots & r_{1p} \\ r_{21}r_{22} & \dots & r_{2p} \\ \dots & \dots & \dots \\ r_{p1}r_{p2} & \dots & r_{pp} \end{pmatrix}, \tag{5.39}$$

où les éléments r_{jk} se déduisent à partir des éléments σ_{jk} à l'aide de la normalisation

$$r_{jk}=\frac{\sigma_{jk}}{(\sigma_{jj}\sigma_{kk})^{1/2}}. \tag{5.40}$$

Les caractéristiques r_{jk} s'appellent *coefficients de corrélation* du couple de variables aléatoires $\xi^{(j)}$ et $\xi^{(k)}$. Elles *mesurent le degré de liaison statistique linéaire* entre ces variables et jouissent des propriétés suivantes:

a) $-1 \leqslant r_{jk} \leqslant 1$; ceci résulte immédiatement des inégalités

$$\mathbf{E}\left(\frac{\xi^{(j)}-m_1^{(j)}}{\sigma_{jj}^{1/2}} \pm \frac{\xi^{(k)}-m_1^{(k)}}{\sigma_{kk}^{1/2}}\right)^2 \geqslant 0;$$

b) la corrélation la plus étroite est réalisée lorsque le coefficient de corrélation est égal à $+1$ ou -1. Dans le premier cas, on a affaire à la corrélation d'une variable aléatoire avec elle-même (il est alors évident que $r_{jj} \equiv 1$); dans le second cas, à la corrélation de deux variables $\xi^{(j)}$ et $\xi^{(k)}$ *linéairement liés*, c'est-à-dire que $\xi^{(j)} = b_0 + b_1\xi^{(k)}$, où b_0 et b_1 sont des constantes (cette liaison est dite *positive* si $b_1 > 0$, *négative* si $b_1 < 0$);

c) si les composantes aléatoires $\xi^{(j)}$ et $\xi^{(k)}$ sont statistiquement indépendantes, alors $r_{jk} = 0$ (ceci résulte directement du fait que pour des variables indépendantes $\xi^{(j)}$ et $\xi^{(k)}$, on a $\mathbf{E}(\xi^{(j)}\cdot\xi^{(k)}) = (\mathbf{E}\xi^{(j)})(\mathbf{E}\xi^{(k)})$. La réciproque (c'est-à-dire que $r_{jk} = 0$ entraîne l'indépendance de $\xi^{(j)}$ et $\xi^{(k)}$) n'est valable que pour certains cas particuliers (par exemple, pour des variables normales $\xi^{(j)}$ et $\xi^{(k)}$). Dans le cas général elle est mise en défaut.

Conclusions

1. Les *variables aléatoires* soumises à un traitement statistique dans le cadre d'expériences aléatoires (ou d'observations) sont de trois types: quantitatives, ordinales et nominales.

2. Les *valeurs possibles* *) de la variable aléatoire dépendent de la nature et de la composition de l'espace fondamental Ω: à toute

*) Outre la forme scalaire, ces « valeurs possibles » peuvent être représentées sous les formes vectorielle et matricielle et même être définies sur un espace de nature très générale (selon le contenu et les objectifs de l'expérience aléatoire effectuée).

issue élémentaire ω est associée une « valeur possible » $x(\omega)$ de la variable aléatoire ξ, donc cette dernière peut être définie comme une fonction $\xi = \xi(\omega)$ définie sur Ω.

3. Les *valeurs observées* $X_1, X_2, \ldots, X_n$ de la variable aléatoire sont des valeurs numériques, vectorielles ou matricielles de ξ réalisées dans n expériences (ou n observations). Le nombre total n des valeurs observées peut être supérieur, égal ou inférieur au nombre total des valeurs théoriquement possibles.

4. La *loi de probabilité* de la variable aléatoire étudiée permet d'associer à tout domaine mesurable ΔX des valeurs possibles la probabilité $p(\Delta X)$ de l'événement « la valeur de ξ réalisée au cours d'une expérience aléatoire (observation) appartient à ce domaine », c'est-à-dire que

$$p(\Delta X) = P\{\xi \in \Delta X\}.$$

5. Pour décrire la loi de probabilité d'une *variable aléatoire multidimensionnelle* $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ on peut comme en dimension un se servir de la fonction de répartition $F_\xi(x^{(1)}, \ldots, x^{(p)}) = P\{\xi^{(1)} < x^{(1)}, \ldots, \xi^{(p)} < x^{(p)}\}$ et de la densité de probabilité $f_\xi(x^{(1)}, \ldots, x^{(p)}) = \partial^p F_\xi(x^{(1)}, \ldots, x^{(p)})/\partial x^{(1)}, \ldots, \partial x^{(p)}$. Mais contrairement à la dimension un, *seule la densité de probabilité définit de façon exhaustive la loi de probabilité multidimensionnelle.*

6. La loi conjointe d'une variable aléatoire multidimensionnelle $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ permet de déduire la *loi marginale* de tout sous-vecteur $\xi_1 = (\xi^{(i_1)}, \xi^{(i_2)}, \ldots, \xi^{(i_k)})$, $1 \leqslant i_q \leqslant p$, $1 \leqslant k \leqslant p$, ainsi que la *loi conditionnelle* de tout sous-vecteur ξ_1 sachant que tout ou partie des autres composantes de ξ sont fixées à des niveaux donnés (cf. (5.15) et (5.16)).

7. Si les composantes $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ de la variable aléatoire $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ sont *statistiquement indépendantes*, la loi de probabilité conjointe $f_\xi(x^{(1)}, \ldots, x^{(p)})$ peut être décrite par p lois marginales à une dimension, puisque, dans ce cas, on a par définition

$$f_\xi(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = f_{\xi^{(1)}}(x^{(1)}) \cdot f_{\xi^{(2)}}(x^{(2)}) \ldots f_{\xi^{(p)}}(x^{(p)}).$$

8. La *population générale* est synonyme d'« espace probabilisé », de « variable aléatoire », de « loi de probabilité » et se définit comme l'ensemble de toutes les observations possibles et imaginables susceptibles d'être effectuées sous le système de conditions donné.

9. L'*échantillon* est une partie de la population générale traitée par les méthodes statistiques pour étudier les propriétés de la population générale. Le problème de la représentativité de l'échantillon implique de l'analyste subtilité et maîtrise des diverses formes d'organisation des sondages pour compenser le manque de temps et de moyens pour obtenir cet échantillon.

10. La connaissance de quelques caractéristiques numériques, notamment, la moyenne $E\xi$, la variance $V\xi$, les coefficients de dissymétrie et d'aplatissement (β_1 et β_2), et dans le cas multidimensionnel, encore les éléments σ_{jk} de la matrice des covariances Σ, suffit pour étudier le comportement de la variable aléatoire ξ.

11. Le calcul des valeurs théoriques des caractéristiques mentionnées implique la connaissance de la densité de probabilité (ou du polygone de fréquence) de la loi analysée. Mais en pratique on les remplace souvent par leurs analogues *empiriques* qui sont déterminés uniquement à l'aide des données d'échantillonnage $X_1, X_2, \ldots, X_n$. Ce changement approximatif est légitimé par le fait que l'échantillon est interprété comme un modèle réduit de la population générale étudiée dans lequel les valeurs observées (c'est-à-dire réalisées) $X_1, X_2, \ldots, X_n$ sont considérées comme possibles, et les probabilités de leur réalisation, prises égales aux fréquences relatives de leur apparition, c'est-à-dire à $1/n$.

12. Une méthode payante d'analyse des propriétés d'une variable aléatoire multidimensionnelle ξ consiste à ranger ses valeurs observées $x_1, x_2, \ldots, x_n$ par ordre de grandeur croissante. Les observations *ordonnées* forment une série $\{x_{(i)}\}_{i=1,\ldots,n}$ appelée *série variationnelle* et ses termes, *statistiques de rang*. Elles sont largement utilisées dans la construction des *estimations non paramétriques* et des *tests non paramétriques* (cf. §§ 8.6, 10.4, 11.2, 11.3).

13. L'échantillon (ou la série variationnelle) est utilisé selon le contexte dans l'*une des deux acceptions* suivantes. Dans la première (l'acception *pratique*), par $X_1, X_2, \ldots, X_n$ on comprend les valeurs de ξ *réellement observées* au cours de l'expérience, c'est-à-dire des nombres ou des vecteurs concrets. Dans la deuxième (l'acception *hypothétique*) $X_1, X_2, \ldots, X_n$ représentent *seulement les notations* des n valeurs (nombres ou vecteurs) *que nous aurions dû obtenir* en effectuant n fois l'expérience (ou l'observation) dans le système de conditions induisant la population générale. Dans le dernier cas, les X_i et toute fonction de X_i n'interviennent plus comme des nombres ou des vecteurs concrets, mais comme des variables aléatoires.

CHAPITRE 6

LES LOIS DE PROBABILITÉ MODÈLES LES PLUS COURAMMENT UTILISÉES DANS LES RECHERCHES STATISTIQUES

Le poids de telle ou telle loi de probabilité dans les recherches statistiques doit être jugé à travers les deux fonctions qu'elle est susceptible de remplir. La première consiste à décrire de façon adéquate le mécanisme du *processus réel* étudié induisant la population générale soumise au traitement statistique. Dans ce cas, *la loi modèle* retenue pour des raisons diverses (ou déduite théoriquement) *décrit la répartition des probabilités d'une variable aléatoire admettant une signification physique bien claire* (salaire d'un travailleur, revenu d'une famille, nombre de pannes d'une chaîne par unité de temps, nombre de pièces défectueuses dans un lot de taille donnée, etc.). Les diverses méthodes de construction, d'analyse et de justification de ces lois relèvent de la modélisation (cf. chap. 3).

L'autre fonction de ces lois modèles est une fonction d'instrument technique auxiliaire dans la réalisation des méthodes de traitement statistique des données. Ces lois permettent de décrire les répartitions des probabilités de certaines fonctions auxiliaires des variables aléatoires étudiées, utilisées pour la construction de toute sorte d'estimations et de tests statistiques (sur les méthodes de construction des estimations et des tests voir §§ 8.1 à 8.6 et 9.1 à 9.6). Signalons en tout premier lieu la répartition du χ^2, la répartition de Student (ou répartition t) et la répartition F.

6.1. Lois de probabilité servant à décrire les mécanismes de processus ou de systèmes réels

6.1.1. Lois décrivant une suite d'épreuves de Bernoulli : la loi binomiale et la loi binomiale négative. Une importante classe de variables aléatoires est induite par une suite d'expériences aléatoires du type suivant : chaque expérience (observation) aléatoire fait apparaître (avec une probabilité p) ou non (respectivement avec une probabilité $q = 1 - p$) l'événement A qui nous intéresse ; *la probabilité p de l'événement A reste la même si cette expérience est répétée m fois, quant aux observations qui constituent cette suite, elles sont mutuellement indépendantes.* Une telle suite d'expériences est appelée *suite d'épreuves de Bernoulli.* On peut décrire cette suite en termes de

variables aléatoires en associant à l'épreuve i la variable aléatoire

$$\xi_i = \begin{cases} 1 \text{ si l'événement } A \text{ se produit;} \\ 0 \text{ sinon.} \end{cases} \tag{6.1}$$

Dans ces conditions, dire qu'une suite $\xi_1, \xi_2, \ldots, \xi_m$ est de Bernoulli revient à dire que $P\{\xi_1 = 1\} = P\{\xi_2 = 1\} = \ldots = P\{\xi_m = 1\} = p$, les variables aléatoires $\xi_1, \xi_2, \ldots, \xi_m$ étant statistiquement indépendantes (la définition de l'indépendance statistique est donnée au n° 5.5.3).

Sous certaines conditions (qui sont en principe approximativement réalisées dans la pratique), la suite d'épreuves de Bernoulli se prête bien à la description d'expériences aléatoires telles que le lancement d'une pièce de monnaie ou d'un dé, le contrôle (d'après le principe binaire) de pièces produites en série, la demande d'un service (avec les issues « libre-occupé »), la réalisation d'une tâche (avec les issues « réalisée-non réalisée »), le tir à la cible (avec les issues « touché-raté »), etc.

Une épreuve de Bernoulli peut être interprétée comme le prélèvement d'un objet dans une population générale infinie dans laquelle p objets possèdent la propriété qui nous intéresse. Dans ce cas, l'événement A consiste en ce que l'objet prélevé possède la propriété indiquée.

La *loi binomiale* décrit la répartition de la variable aléatoire $\nu_p(m) = \xi_1 + \xi_2 + \ldots + \xi_m$, c'est-à-dire du nombre d'apparitions d'un événement dans une suite de m épreuves indépendantes, la probabilité d'apparition de cet événement dans une épreuve étant égale à p.

De la définition de la variable aléatoire binomiale, il s'ensuit que ses valeurs possibles sont tous les entiers positifs de 0 à m. Pour calculer les probabilités $P\{\nu_p(m) = x\}$ ($x = 0, 1, 2, \ldots, m$), étudions attentivement l'espace fondamental engendré par la suite d'épreuves de Bernoulli. Il est évident qu'à tout événement élémentaire ω est associée une suite de 0 et de 1 de longueur m, soit

$$\delta_1(\omega), \delta_2(\omega), \ldots, \delta_m(\omega). \tag{6.2}$$

Partageons ces suites en classes en rapportant à la classe d'indice x toutes les suites de type (6.2) contenant un *même* nombre x d'unités :

$$\begin{array}{ll} x = 0: & (0, 0, \ldots, 0, 0) = \omega(0) \\ x = 1: & \left\{ \begin{array}{l} (1, 0, \ldots, 0, 0) = \omega_1(1) \\ (0, 1, \ldots, 0, 0) = \omega_2(1) \\ \ldots\ldots\ldots\ldots\ldots \\ (0, 0, \ldots, 0, 1) = \omega_m(1) \end{array} \right. \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}$$

$$
\begin{array}{ll}
x=k: & \left\{\begin{array}{l}
\overbrace{(1,\ 1,\ 1,\ \ldots,\ 1,}^{k}\ 0\ \ldots,\ 0,\ 0)=\omega_1(k) \\
(0,\ 1,\ \ldots,\ 0,\ \ldots,\ 0,\ 0)=\omega_2(k) \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
(0,\ 0,\ \ldots,\ 0,\ \ldots,\ 1,\ 1)=\omega_{N(x)}(k)
\end{array}\right. \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
k=m: & (1,\ 1,\ \ldots,\ 1,\ \ldots,\ 1,\ 1)=\omega(m).
\end{array}
$$

Etant donné que le nombre $N(x)$ d'événements élémentaires de la classe d'indice x est égal à C_m^x (le nombre de combinaisons de x unités m à m) et que la probabilité de réalisation de toute issue élémentaire de la classe d'indice x est égale visiblement à $p^x(1-p)^{m-x}$, on obtient

$$
P\{\nu_p(m)=x\}=P\{\omega_1(x)+\omega_2(x)+\ldots+\omega_{C_m^x}(x)\}=
$$

$$
=\sum_1^{C_m^x} p^x(1-p)^{m-x}=C_m^x p^x(1-p)^{m-x}. \qquad (6.3)
$$

Ceci est l'expression analytique de la loi binomiale *). Ses principales caractéristiques numériques (qui ici se calculent plus facilement non pas à l'aide des formules directes (5.21), mais en utilisant la relation $\nu_p(m)=\xi_1+\ldots+\xi_m$, l'indépendance mutuelle des ξ_i et la simplicité des moments) sont:

la *moyenne*: $E\nu_p(m)=mp$;

le *mode*: $p(m+1)-1\leqslant x_{\text{mod}}\leqslant p(m+1)$;

la *variance*: $V\nu_p(m)=mp(1-p)$;

le *coefficient de dissymétrie* : $\beta_1=\dfrac{1-2p}{\sqrt{mp(1-p)}}$;

le *coefficient d'aplatissement*: $\beta_2=\dfrac{1-6p(1-p)}{mp(1-p)}$.

La loi binomiale est largement utilisée pour le contrôle de qualité des produits, pour la description du fonctionnement des systèmes de files d'attente, en théorie du tir et dans d'autres domaines des activités pratiques.

La *loi binomiale négative* décrit la répartition d'une variable aléatoire $\nu_p^-(k)$ définie par une suite d'épreuves de Bernoulli (cf.

*) Cette loi s'appelle binomiale, car le second membre de (6.3) est le x-ième terme du développement du binôme $(p+(1-p))^m$. Ce même fait permet de vérifier l'axiome de normalisation des probabilités, c'est-à-dire la véracité de l'identité $\sum_{x=0}^{m} C_m^x p^x(1-p)^{m-x}=1$.

(6.1)) de la manière suivante:

$$\sum_{i=1}^{\nu_p^-(k)-1} \xi_i = k-1, \qquad \sum_{i=1}^{\nu_p^-(k)} \xi_i = k.$$

En d'autres termes, $\nu_p^-(k)$ est le nombre d'épreuves nécessaire pour faire apparaître dans une suite de Bernoulli k fois un événement (dont la probabilité d'apparition au cours d'une épreuve est égale à p). Le calcul de l'expression analytique de la loi de probabilité de la variable aléatoire $\nu_p^-(k)$ est facile. Figeons une valeur possible quelconque x de $\nu_p^-(k)$. Du fait que l'événement qui nous intéresse s'est réalisé k fois en $\nu_p^-(k) = x$ épreuves, il s'ensuit qu'il s'est réalisé $k-1$ fois en $x-1$ épreuves. Le théorème de multiplication des probabilités nous donne alors:

$$P\{\nu_p^-(k) = x\} = [C_{x-1}^{k-1} p^{k-1} (1-p)^{(x-1)-(k-1)}] \cdot p =$$
$$= C_{x-1}^{k-1} p^k (1-p)^{x-k}, \quad x = k,\ k+1,\ \ldots \qquad (6.4)$$

Cette loi tire son nom de ce que les seconds membres de (6.4) sont les termes du *binôme d'exposants négatifs*: $p^k (1-(1-p))^{-k}$.

Les principales caractéristiques numériques de cette loi sont:

la *moyenne*: $\mathrm{E}\nu_p^-(k) = \dfrac{k}{p}$;

la *variance*: $\mathrm{V}\nu_p^-(k) = \dfrac{k(1-p)}{p^2}$;

le *coefficient de dissymétrie*: $\beta_1 = \dfrac{2-p}{\sqrt{k(1-p)}}$;

le *coefficient d'aplatissement*: $\beta_2 = \dfrac{1+4(1-p)+(1-p)^2}{k(1-p)}$.

La loi binomiale négative est appliquée à la statistique des accidents et des maladies, dans les problèmes d'analyse des quantités d'individus d'une espèce donnée contenus dans des échantillons biologiques, dans les problèmes de réservation optimale d'éléments, en théorie du tir.

6.1.2. Loi hypergéométrique. Dans l'une des interprétations de la variable aléatoire binomiale $\nu_p(m)$ nous avons considéré une population générale infinie dont p objets possédaient le caractère qui nous intéresse. Alors $\nu_p(m)$ représente le nombre d'objets possédant ce caractère parmi m objets tirés au hasard dans la population générale. La *variable aléatoire hypergéométrique* $\nu_{MN}(m)$ peut être traitée comme une modification de la variable aléatoire binomiale $\nu_p(m)$ rapportée à une population générale *finie* constituée de N objets dont M possèdent le caractère étudié. Autrement dit, $\nu_{MN}(m)$ est le nombre d'objets jouissant de ce caractère parmi m objets tirés (sans remise) au hasard dans une population de N objets dont M

ont ce caractère. Il est évident que les valeurs possibles de la variable aléatoire $\nu_{M,N}(m)$ seront tous les entiers positifs de max $\{0, m - (N - M)\}$ à min $\{m, M\}$. Pour établir la forme analytique de la loi de probabilité, calculons la probabilité de l'événement $\{\nu_{M,N}(m) = x\}$ comme le rapport du nombre de tous les échantillons de taille m conduisant à la réalisation de cet événement (le nombre de cas favorables) au nombre de façons de choisir m objets parmi N (le nombre de toutes les issues possibles). Il est évident qu'à toute collection de x objets possédant le caractère étudié sont associées C_{N-M}^{m-x} façons de choisir les $m - x$ objets restants parmi les objets ne possédant pas ce caractère. Comme le nombre de collections de x objets jouissant de ce caractère est égal à C_M^x, le nombre total d'issues favorables (de l'événement $\{\nu_{M,N}(m) = x\}$) sera égal à $C_M^x \times C_{N-M}^{m-x}$. Vu que le nombre de toutes les issues possibles, c'est-à-dire de toutes les façons dont on peut extraire m objets parmi N, est égal à C_N^m, il vient

$$P\{\nu_{M,N}(m) = x\} = \frac{C_M^x \cdot C_{N-M}^{m-x}}{C_N^m}. \tag{6.5}$$

Cette loi est souvent utilisée pour le contrôle de produits industriels, ainsi que dans les problèmes d'organisation de sondages.

Ses principales caractéristiques sont:

la *moyenne*: $\mathrm{E}\nu_{M,N}(m) = m\dfrac{M}{N}$;

la *variance*: $\mathrm{V}\nu_{M,N}(m) = m\dfrac{M}{N-1}\left(1-\dfrac{M}{N}\right)\left(1-\dfrac{m}{N}\right)$;

le *coefficient de dissymétrie*: $\beta_1 =$

$$= \frac{\left(1-2\cdot\dfrac{M}{N}\right)}{\sqrt{m\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)}}\;\frac{(N-2m)\sqrt{N-1}}{(N-2)\sqrt{N-m}};$$

le *coefficient d'aplatissement*: $\beta_2 =$

$$= \frac{c_1(N) - c_2(N)\cdot 6\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)}{m\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)} + c_3(N) + c_4(N),$$

où

$$c_1(N) = \frac{(N-1)N(N+1)}{(N-2)(N-3)(N-m)};$$

$$c_2(N) = \frac{(N-1)N^2}{(N-2)(N-3)(N-m)};$$

$$c_3(N) = 3\left[\frac{(N-1)N^2}{(N-2)(N-3)(N-m)} - 1\right];$$

$$c_4(N) = \frac{18(N-1)}{(N-2)(N-3)} - \frac{6(N-1)}{(N-2)(N-3)m\frac{M}{N}\left(1-\frac{M}{N}\right)} - \frac{3(N-1)Nm}{(N-2)(N-3)(N-m)}.$$

Lorsque $N \to \infty$, le second membre de (6.5) tend en toute logique vers l'expression (6.3) de la loi binomiale. La moyenne, la variance, les coefficients de dissymétrie et d'aplatissement de la loi hypergéométrique tendent respectivement vers les mêmes caractéristiques numériques de la loi binomiale (ceci s'établit sans peine par un passage à la limite).

6.1.3. Loi de Poisson. La *loi de Poisson* (qui a été publiée par ce savant en 1837) décrit le nombre d'apparitions pendant une unité de temps d'un événement dont la réalisation ne dépend pas du nombre de réalisations passées et n'influe pas sur les futures, les épreuves se déroulant dans des conditions stationnaires. Cette loi peut également être décrite comme la *limite de la loi binomiale* lorsque la probabilité p d'apparitions de l'événement étudié dans une épreuve est très petite et le nombre m d'épreuves effectuées en une unité de temps, assez élevé, plus exactement lorsque le produit $mp \to \lambda$ ($\lambda = \text{const} > 0$) pour $p \to 0$ et $m \to \infty$. La loi de Poisson est pour cette raison appelée aussi *loi des événements rares.* Désignons par $\nu_0(\infty)$ ou simplement ν_0 une variable aléatoire poissonnienne (cette notation exprime que $\nu_0(\infty)$ est déduite de la variable aléatoire binomiale $\nu_p(m)$ lorsque $p \to 0$ et $m \to \infty$) et calculons sa loi de probabilité

$$P\{\nu_0 = x\} = \lim_{\substack{p\to 0,\ m\to\infty \\ (pm\to\lambda)}} P\{\nu_p(m) = x\} = \lim C_m^x p^x (1-p)^{m-x} =$$

$$= \lim \frac{m(m-1)\ldots(m-x+1)}{x!}\,\frac{\lambda^x}{m^x}\cdot\left(1-\frac{\lambda}{m}\right)^{m-x} =$$

$$= \lim\left[\frac{\lambda^x}{x!}\cdot\left(1-\frac{\lambda}{m}\right)^m\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\ldots\right.$$

$$\left.\ldots\left(1-\frac{x-1}{m}\right)\left(1-\frac{\lambda}{m}\right)^{-x}\right] = \frac{\lambda^x}{x!}\cdot\lim_{m\to\infty}\left(1-\frac{\lambda}{m}\right)^m \times$$

$$\times \lim_{m\to\infty}\left[\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\ldots\left(1-\frac{x-1}{m}\right)\left(1-\frac{\lambda}{m}\right)^{-x}\right] =$$

$$= \frac{\lambda^x}{x!}\,e^{-\lambda}\ (x = 0,\ 1,\ 2\ldots). \quad (6.6)$$

On voit donc que la loi de Poisson dépend de l'unique paramètre λ qui figure le nombre moyen de réalisations de l'événement étudié pendant une unité de temps.

Les formules (5.20) et (5.21) nous permettent de calculer directement les caractéristiques numériques de la loi de Poisson:

la *moyenne*: $\mathrm{E}\nu_0 = \lambda$;

la *variance*: $\mathrm{V}\nu_0 = \lambda$;

le *coefficient de dissymétrie*: $\beta_1 = \frac{1}{\sqrt{\lambda}}$;

le *coefficient d'aplatissement*: $\beta_2 = \frac{1}{\lambda}$.

La variable aléatoire poissonnienne décrit le nombre de pannes d'une chaîne ou d'un système complexe (fonctionnant en régime « normal ») par unité de temps; le nombre d'arrivées par unité de temps dans une station; les lois statistiques des sinistres et des maladies rares.

Cette loi n'est pas séduisante uniquement par le calcul élémentaire de ses caractéristiques numériques et le laconisme de la formule (6.6) (qui met en jeu un *seul paramètre* λ!). Cette loi est performante dans les situations qui s'écartent de la procédure de sa formation. On peut, par exemple, admettre que dans les épreuves de Bernoulli, l'événement étudié se réalise avec des probabilités distinctes $p_1, p_2, \ldots, p_n$. Dans une telle situation, la loi binomiale ne passe pas contrairement à l'expression (6.6) qui reste valable approximativement et nous donne une description assez exacte de la répartition de la variable aléatoire étudiée, pourvu qu'on y remplace $\lambda = np$ par $\lambda = n\bar{p}$, où $\bar{p} = (p_1 + \ldots + p_n)/n$. D'après ce qui précède, on peut supposer que la population analysée est composée d'un amalgame de sous-populations telles que le nombre p d'objets possédant le caractère requis varie d'une sous-population à l'autre, donc, le nombre moyen λ de réalisations en une unité de temps de l'événement étudié. On peut prouver par ailleurs que si l'on étudie non pas la valeur moyenne de ces p (ou λ), mais le paramètre en le traitant comme une *variable aléatoire*, on sera conduit à une loi qui dans un certain sens est proche de la loi de Poisson. Si, par exemple, on admet que la fonction de répartition du paramètre λ est de la forme

$$f_\lambda(x) = \frac{p^k}{\Gamma(k)} \cdot x^{k-1} e^{-px},$$

où $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx$ est la fonction gamma; $k > 0$ et $p > 0$ les paramètres de la loi; $x > 0$ les valeurs possibles de λ; le nombre de réalisations (en une unité de temps) de l'événement étudié suivra la loi binomiale négative (6.4) (pour plus de détails sur la répartition $f_\lambda(x)$, cf. n° 6.2.5).

6.1.4. Loi polynomiale. La population générale polynomiale est une généralisation utile de la population binomiale au cas de

plus de deux issues possibles. La population générale polynomiale est infinie et contient des objets de classes 1, 2, . . ., l ($l > 2$) respectivement en quantités $p_1, p_2, \ldots, p_l$ (dans la population binomiale on avait $l = 2$, $p_1 = p$ et $p_2 = 1 - p$). Donc, un objet de classe j apparaît avec la probabilité p_j dans une expérience aléatoire (un tirage aléatoire dans la population générale). On se propose d'étudier la loi de probabilité d'une variable aléatoire ($\nu_p^{(1)}(m)$, $\nu_p^{(2)}(m)$, . . ., $\nu_p^{(l)}(m)$) engendrée par une expérience répétée m fois (c'est-à-dire par un échantillon de m objets), où $\nu_p^{(j)}(m)$ est le nombre d'objets de la classe j contenus dans cet échantillon, $p = (p_1, p_2, \ldots, p_l)$ (il est évident que $\sum_{j=1}^{l} p_j = 1$ et $\sum_{j=1}^{l} \nu_p^{(j)}(m) = m$).

La répartition discrète multidimensionnelle correspondante est définie par l'expression (qui s'établit par des raisonnements probabilistes directs)

$$P\{\nu_p^{(1)}(m) = x^{(1)}, \ldots, \nu_p^{(l)}(m) = x^{(l)}\} = \\ = \frac{m!}{x^{(1)}!\,x^{(2)}!\ldots x^{(l)}!}\, p_1^{x^{(1)}} \cdot \ldots \cdot p_l^{x^{(l)}}, \qquad (6.7)$$

où $x^{(1)}, x^{(2)}, \ldots, x^{(l)}$ sont des nombres entiers positifs quelconques (donnés) vérifiant la condition $\sum_{j=1}^{l} x^{(j)} = m$. L'expression (6.7) définit la probabilité que parmi les m objets tirés, $x^{(1)}$ sont de classe 1, $x^{(2)}$ de classe 2, etc. On peut aussi rattacher la variable aléatoire polynomiale à une expérience aléatoire répétée m fois, donnant lieu à chaque fois à l'une des l issues possibles $A_1, A_2, \ldots, A_l$, la probabilité de l'issue A_j étant p_j.

Cette loi doit son nom au fait que l'expression (6.7) est le terme général du *polynôme* $(p_1 + p_2 + \ldots + p_l)^m$.

Le vecteur des moyennes ($\mathrm{E}\nu_p^{(1)}(m), \ldots, \mathrm{E}\nu_p^{(l)}(m)$) et les covariances $\sigma_{jk} = \mathrm{E}\{(\nu_p^{(j)}(m) - \mathrm{E}\nu_p^{(j)}(m))(\nu_p^{(k)}(m) - \mathrm{E}\nu_p^{(k)}(m))\}$ des composantes de la variable aléatoire vectorielle étudiée sont définies par les expressions:

les *moyennes*: $\mathrm{E}\nu_p^{(j)}(m) = mp_j$; $j = 1, 2, \ldots, l$;

les *variances*: $\mathrm{V}\nu_p^{(j)}(m) = \sigma_{jj} = mp_j(1 - p_j)$, $j = 1, 2, \ldots, l$;

les *covariances*: $\sigma_{jk} = -mp_jp_k$; $j, k = 1, 2, \ldots, l$, $j \neq k$.

La loi polynomiale est appliquée essentiellement au traitement statistique des échantillons de grosses populations dont les éléments se répartissent en plus de deux catégories (par exemple, dans les sondages sociologiques, sociologico-économiques, médicaux et autres).

6.1.5. Loi normale (de Gauss). Cette loi est capitale dans la théorie et la pratique des recherches probabilisto-statistiques. En

1733 elle fut étudiée par Moivre pour la première fois en tant qu'approximation continue de la loi binomiale (voir plus bas le théorème de Moivre-Laplace, § 7.3). La loi normale fut redécouverte plus tard et étudiée indépendamment l'un de l'autre par K. Gauss (1809) et P. Laplace (1812). Ces deux savants sont arrivés à la fonction de répartition normale dans le cadre de leurs travaux sur la théorie des erreurs d'observation. Leur idée du mécanisme de formation des variables aléatoires normales est la suivante. On postule que les valeurs prises par la variable aléatoire continue étudiée *dépendent d'un très grand nombre de facteurs aléatoires indépendants, l'action de chacun d'eux étant très petite et ne prédominant pas celles des autres et le caractère de cette action étant additif* (autrement dit, si un facteur aléatoire F agit sur la quantité a, on obtient la quantité $a + \Delta(F)$, où la « contribution » aléatoire $\Delta(F)$ est petite et de signe équiprobable *). On démontre que la densité de probabilité des variables aléatoires de ce type est de la forme

$$\varphi(x;\ a,\ \sigma^2) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\, e^{-\frac{(x-a)^2}{2\sigma^2}}, \tag{6.8}$$

où a et σ^2, les paramètres de la loi, figurent respectivement la moyenne et la variance de la variable aléatoire (à tout seigneur tout honneur, la densité de probabilité et la fonction de répartition de la loi normale seront désignées par des notations spéciales en raison du rôle particulier qu'elles jouent).

La fonction de répartition de la variable aléatoire normale $\xi(a,\ \sigma^2)$ sera notée par $\Phi(x;\ a,\ \sigma^2)$ et définie par l'expression

$$\Phi(x;\ a,\ \sigma^2) = P\{\xi(a,\ \sigma^2) < x\} = \frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}}\, dt. \tag{6.8'}$$

Convenons d'appeler *réduite* la loi normale de paramètres $a = 0$ et $\sigma^2 = 1$ et de désigner sa densité de probabilité et sa fonction de répartition respectivement par $\varphi(x) = \varphi(x;\ 0,\ 1)$ et $\Phi(x) = \Phi(x;\ 0,\ 1)$.

Dans de nombreuses variables aléatoires étudiées en économie, en technique, en médecine, en biologie et dans d'autres domaines, il est naturel de voir l'effet conjugué d'un grand nombre de facteurs indépendants. Mais la place privilégiée de la loi normale ne doit pas être expliquée par son universalité comme on le pensait depuis bien longtemps (probablement sous l'influence des brillants travaux de Gauss et Laplace). De ce point de vue, la loi normale est une loi comme une autre, avec, il est vrai, le plus vaste champ d'applica-

*) Ces conditions sont rigoureusement formalisées, par exemple, dans le théorème limite central (cf. § 7.3).

tions. Et l'on comprend la pointe d'ironie lancée par Lipman (et rapportée par H. Poincaré dans son « Calcul des probabilités », Paris, 1912) : « Chacun est convaincu de la véracité de la loi normale : les expérimentateurs, parce qu'ils pensent que c'est un théorème de mathématique ; les mathématiciens, parce qu'ils pensent que c'est un fait expérimental ». Il n'en reste pas moins que la loi normale est la plus attrayante et la plus commode pour la relative simplicité de ses propriétés et parce que tout est connu sur elle sur le plan théorique. Même si les données empiriques s'écartent de la loi normale, il existe pour le moins deux façons de la faire intervenir : a) l'utiliser comme première approximation ; il n'est pas rare que les résultats obtenus sous cette condition soient satisfaisants pour les objectifs poursuivis ; b) transformer la variable aléatoire ξ de telle sorte que la loi initiale devienne normale. La loi normale est commode pour les applications par sa propriété d'« autoreproduction » qui se traduit par le fait que la somme de variables aléatoires normales suit la loi normale. Par ailleurs, la loi normale revêt une importante signification théorique : elle a permis d'établir d'importantes lois et de construire divers tests statistiques, etc. (loi du χ^2, loi de Student, loi F et les tests respectifs, cf. n° 6.2.1, 6.2.2 et 6.2.3, ainsi que le chapitre 11).

Les graphiques des densités normales sont représentés sur les figures 5.5, 5.6, 5.10 et 5.11.

Les principales caractéristiques numériques de la loi normale sont :

la *moyenne, mode, médiane* : $\mathbf{E}\xi = x_{\text{mod}} = x_{\text{méd}} = a$;

la *variance* : $\mathbf{V}\xi = \sigma^2$;

le *coefficient de dissymétrie* : $\beta_1 = 0$;

le *coefficient d'aplatissement* : $\beta_2 = 0$.

La **loi normale bidimensionnelle** décrit la répartition conjointe d'une variable aléatoire bidimensionnelle $\xi = (\xi^{(1)}, \xi^{(2)})$ de composantes $\xi^{(1)}$ et $\xi^{(2)}$ continues et dont le mécanisme de formation est le même qu'en dimension un, les ensembles de facteurs aléatoires agissant sur $\xi^{(1)}$ et $\xi^{(2)}$ étant généralement non disjoints (d'où la dépendance éventuelle de $\xi^{(1)}$ et $\xi^{(2)}$).

Les principales caractéristiques numériques de la variable aléatoire bidimensionnelle $\xi = (\xi^{(1)}, \xi^{(2)})$ sont :

le *vecteur des moyennes* : $M_1 = \begin{pmatrix} m_1^{(1)} \\ m_1^{(2)} \end{pmatrix}$, où $m_1^{(i)} = \mathbf{E}\xi^{(i)}$;

la *matrice des covariances* : $\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, où $\sigma_{jk} = \mathbf{E}\{(\xi^{(j)} - m_1^{(j)})(\xi^{(k)} - m_1^{(k)})\}$;

le *coefficient de corrélation* : $r = \dfrac{\sigma_{12}}{(\sigma_{11}\sigma_{22})^{1/2}}$.

La densité conjointe bidimensionnelle $\varphi(x^{(1)}, x^{(2)}) = f_\xi(x^{(1)}, x^{(2)})$ de la loi normale peut être mise sous la forme

$$\varphi(x^{(1)}, x^{(2)}) = \frac{1}{2\pi[\sigma_{11}\sigma_{22}(1-r^2)]^{1/2}} \cdot e^{-\frac{1}{2(1-r^2)}\left[\frac{(x^{(1)}-m_1^{(1)})^2}{\sigma_{11}} - 2r\frac{x^{(1)}-m_1^{(1)}}{\sigma_{11}^{1/2}} \cdot \frac{x^{(2)}-m_1^{(2)}}{\sigma_{22}^{1/2}} + \frac{(x^{(2)}-m_1^{(2)})^2}{\sigma_{22}}\right]} \quad (6.9)$$

ou encore

$$\varphi(x^{(1)}, x^{(2)}) = \varphi(X) = \frac{1}{2\pi \cdot |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(X-M_1)'\Sigma^{-1}(X-M_1)}, \quad (6.9')$$

où $X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$, l'indice « prime » symbolise la transposition d'une matrice ou d'un vecteur, $|\Sigma| = \det(\Sigma)$ est le déterminant de la matrice des covariances, Σ^{-1} l'inverse de la matrice des covariances. La surface de la densité de la loi normale bidimensionnelle est représentée sur la figure 5.7.

Les *densités marginales* $\varphi_{\xi^{(1)}}(x^{(1)})$ et $\varphi_{\xi^{(2)}}(x^{(2)})$ peuvent être déduites de la densité conjointe à l'aide de la formule (5.15):

$$\varphi_{\xi^{(1)}}(x^{(1)}) = \frac{1}{\sqrt{2\pi\sigma_{11}}} e^{-\frac{(x^{(1)}-m_1^{(1)})^2}{2\sigma_{11}}};$$

$$\varphi_{\xi^{(2)}}(x^{(2)}) = \frac{1}{\sqrt{2\pi\sigma_{22}}} e^{\frac{-(x^{(2)}-m_1^{(2)})^2}{2\sigma_{22}}}.$$

Ces formules expriment que les lois marginales des composantes de la variable aléatoire normale bidimensionnelle sont elles-mêmes des lois normales à une dimension respectivement de paramètres $(m_1^{(1)}, \sigma_{11})$ et $(m_1^{(2)}, \sigma_{22})$.

Les *densités conditionnelles* $\varphi_{\xi^{(1)}}((x^{(1)} \mid \xi^{(2)} = x^{(2)})$ et $\varphi_{\xi^{(2)}} \times (x^{(2)} \mid \xi^{(1)} = x^{(1)})$ se calculent à l'aide des formules (5.16) et (5.16'):

$$\varphi_{\xi^{(1)}}(x^{(1)} \mid \xi^{(2)} = x^{(2)}) = \frac{1}{\sqrt{2\pi\sigma_{11}(1-r^2)}} \times e^{-\frac{1}{2\sigma_{11}(1-r^2)}\left[x^{(1)}-\left(m_1^{(1)}+r\frac{\sigma_{11}^{1/2}}{\sigma_{22}^{1/2}}(x^{(2)}-m_1^{(2)})\right)\right]^2};$$

$$\varphi_{\xi^{(2)}}(x^{(2)} \mid \xi^{(1)} = x^{(1)}) = \frac{1}{\sqrt{2\pi\sigma_{22}(1-r^2)}} \times e^{-\frac{1}{2\sigma_{22}(1-r)^2}\left[x^{(2)}-\left(m_1^{(2)}+r\frac{\sigma_{22}^{1/2}}{\sigma_{11}^{1/2}}\cdot(x^{(1)}-m_1^{(1)})\right)\right]^2}.$$

De là on déduit, en particulier, que la loi conditionnelle de la composante $\xi^{(i)}$ sachant que $\xi^{(j)} = x^{(j)}$ est de nouveau une loi normale dont la moyenne dépend, comme il fallait s'y attendre, de la valeur fixée $x^{(j)}$:

$$\mathrm{E}\,(\xi^{(i)} | \xi^{(j)} = x^{(j)}) = m_1^{(j)} + r\frac{\sigma_{ii}}{\sigma_{jj}}(x^{(j)} - m_1^{(j)}),$$

et dont la variance ne dépend pas de $x^{(j)}$ et vaut

$$\mathrm{V}\,(\xi^{(i)} \mid \xi^{(j)} = x^{(j)}) = \sigma_{ii}\,(1 - r^2).$$

La loi normale conjointe décrit la répartition d'une variable aléatoire p-dimensionnelle $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ à composantes continues $\xi^{(j)}$ et dont le mécanisme de formation est pour chacune d'elles le même qu'en dimension un, les ensembles des facteurs aléatoires agissant sur elles étant généralement non disjoints (d'où leur éventuelle dépendance mutuelle). En se donnant le vecteur colonne p-dimensionnel M_1 des moyennes des composantes et la $(p \times p)$-matrice des covariances Σ (cf. n° 5.6.7), on peut exprimer la densité conjointe p-dimensionnelle de la loi normale conjointe sous la forme

$$\varphi\,(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \frac{1}{(2\pi)^{\frac{p}{2}} \,|\, \Sigma \,|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(X-M_1)'\ \Sigma^{-1}\ (X-M_1)}, \qquad (6.10)$$

où comme toujours $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ est le vecteur colonne des variables courantes, $|\,\Sigma\,| = \det(\Sigma)$, le déterminant de la matrice des covariances.

La dégénérescence de la matrice Σ (c'est-à-dire la nullité de son déterminant $|\,\Sigma\,|$) rend la loi correspondante *dégénérée* (ou *singulière*); cela signifie en particulier que les valeurs de la variable aléatoire sont éparpillées dans un sous-espace de dimension inférieure à p. Hormis quelques cas spéciaux, nous admettrons toujours que nous nous trouvons dans cet espace de dimension inférieure à p, de sorte que $|\,\Sigma\,| > 0$.

6.1.6. Loi lognormale. Une variable aléatoire η suit la loi lognormale si son logarithme népérien $\ln \eta$ est une variable normale. Cela signifie en particulier que les valeurs de la variable lognormale dépendent d'un très grand nombre de facteurs mutuellement indépendants, l'action de chacun d'eux étant « uniformément insignifiante » et de signe équiprobable. A noter que contrairement au mécanisme de formation de la loi normale, le caractère séquentiel de l'action des facteurs aléatoires est tel que l'accroissement aléatoire provoqué par l'action d'un facteur est proportionnel à la valeur déjà prise par la variable aléatoire au moment de cette action (on dit alors que l'action des facteurs est *multiplicative*). Traduit dans le

langage mathématique, ceci devient : si $\eta_0 = a$ est une composante *non aléatoire* de la variable η (en quelque sorte la vraie valeur de η dans un schéma idéal soustrait à l'influence de facteurs aléatoires), et $\xi_1, \xi_2, \ldots, \xi_N$ l'expression numérique des effets de l'action des facteurs aléatoires indiqués, alors les valeurs perturbées de la variable seront

$$\eta_1 = \eta_0 + \xi_1 \cdot \eta_0;$$
$$\eta_2 = \eta_1 + \xi_2 \cdot \eta_1;$$
$$\ldots\ldots\ldots\ldots$$
$$\eta_N = \eta_{N-1} + \xi_N \cdot \eta_{N-1}.$$

D'où l'on déduit immédiatement que

$$\sum_{i=0}^{N-1} \left(\frac{\Delta\eta_i}{\eta_i}\right) = \xi_1 + \xi_2 + \ldots + \xi_N, \tag{6.11}$$

où $\Delta\eta_i = \eta_{i+1} - \eta_i$. Mais le second membre de (6.11) est le résultat de l'action *additive* de l'ensemble des facteurs aléatoires, donc d'après les conditions posées ci-dessus, cette somme suit une loi normale (cf. n° 6.1.5, ainsi que le § 7.3 consacré au théorème limite central). Par ailleurs, vu que le nombre des facteurs aléatoires est élevé (c'est-à-dire que $N \to \infty$) et que l'action de chacun d'eux est relativement petite (c'est-à-dire que $\Delta\eta_i \to 0$), on peut dans le premier membre de (6.11) passer à l'intégrale

$$\int_{\eta_0}^{\eta} \frac{d\eta}{\eta} = \ln \eta - \ln \eta_0 = \ln \eta - \ln a.$$

Ce qui signifie en définitive que le logarithme de η (diminué d'une quantité constante $\ln a$) suit une loi normale avec une moyenne nulle, c'est-à-dire que

$$F_\eta(x) = P\{\eta < x\} = P\{\ln \eta < \ln x\} = \frac{1}{\sqrt{2\pi}\cdot\sigma} \int_0^{\ln x} e^{-\frac{(t-\ln a)}{2\sigma^2}}\, dt,$$

d'où l'on déduit, en dérivant les deux membres par rapport à x, que

$$f_\eta(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma x} e^{\frac{-(\ln x - \ln a)^2}{2\sigma^2}} \tag{6.12}$$

(la possibilité d'utiliser l'identité $P\{\eta < x\} = P\{\ln \eta < \ln x\}$ pour le calcul de $f_\eta(x)$ résulte de la stricte monotonie de $\ln \eta$).

Le schéma de formation des valeurs de la variable lognormale qui vient d'être décrit est caractéristique à de nombreuses situations physiques et socio-économiques (dimensions et poids de particules

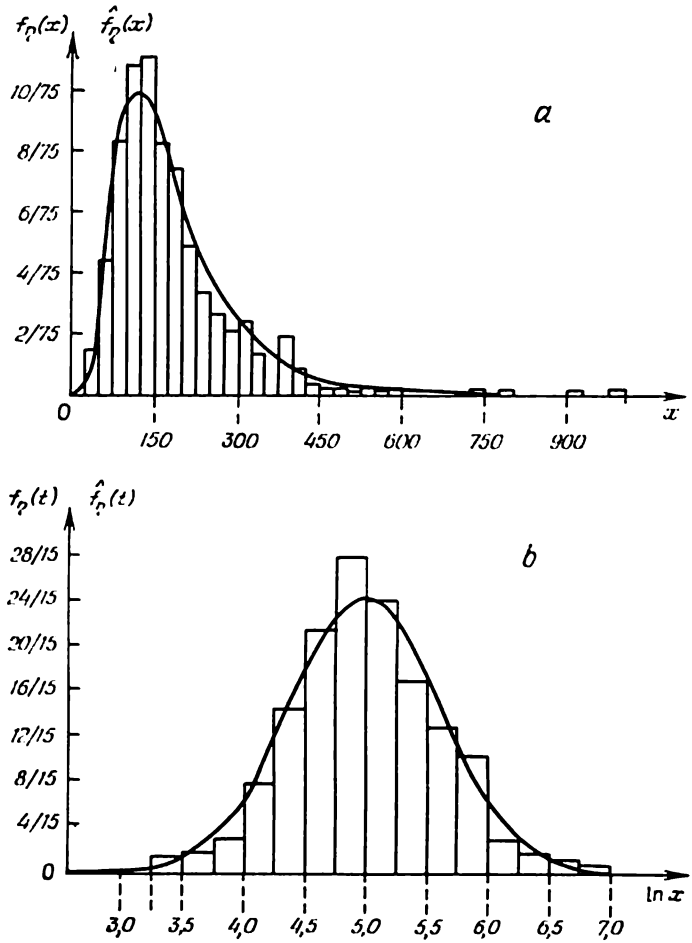


Fig. 6.1. Histogramme et densité théorique (modèle) de la répartition des foyers suivant le revenu mensuel moyen par membre (*a*) et suivant le logarithme du revenu mensuel moyen par membre (*b*)

obtenues par désagrégation; salaire d'un employé; revenu d'une famille; dimensions des formations cosmiques; durée de vie d'un article fonctionnant en régime d'usure et de vieillissement, etc.; cf.; par exemple [2], [3], [4]).

Exemple 6.1.*). La variable aléatoire η représente le revenu mensuel par personne (en dollars) d'une famille dans une population

*) Les données de cet exemple ont été empruntées à *Introduction to Frequency Curves*. Bull. Amer. teleph. telegr. Company, 1953, n° 1.

Tableau 6.1

Numéro de l'intervalle	Nombre de données empiriques tombant dans cet intervalle	Numéro de l'intervalle	Nombre de données empiriques tombant dans cet intervalle
1	2	21	1
2	15	22	3
3	44	23	1
4	83	24	2
5	108	25	0
6	110	26	1
7	83	27	0
8	75	28	1
9	49	29	1
10	34	30	2
11	27	31	0
12	21	32	1
13	24	33	0
14	13	34	0
15	13	35	0
16	19	36	0
17	8	37	1
18	3	38	0
19	2	39	0
20	2	40	1

de $n = 750$ familles. Les tableaux 6.1 et 6.2 représentent respectivement les résultats du groupement des données empiriques (x_i) et leurs logarithmes ($\ln x_i$) (la longueur d'un intervalle de groupement équivaut à 25 dollars). Sur la figure 6.1, *a*, *b* sont tracés les histogrammes et les densités respectivement de la loi lognormale et de la loi normale.

Tableau 6.2

Numéro de l'intervalle	Nombre de données empiriques tombant dans cet intervalle	Numéro de l'intervalle	Nombre de données empiriques tombant dans cet intervalle
1	1	10	141
2	1	11	123
3	0	12	86
4	7	13	65
5	8	14	52
6	15	15	14
7	39	16	8
8	73	17	5
9	109	18	3

Plus bas sont exhibés les résultats du calcul des principales caractéristiques numériques de la loi lognormale (en fonction des paramètres a et σ^2):

la *moyenne* : $\mathrm{E}\eta = a e^{\frac{1}{2}\sigma^2}$;

le *mode* : $x_{\text{mod}} = a e^{-\sigma^2}$;

la *médiane*: $x_{\text{méd}} = a$;

la *variance*: $\mathrm{V}\eta = (\mathrm{E}\eta)^2 (e^{\sigma^2} - 1) = a^2 e^{\sigma^2} (e^{\sigma^2} - 1)$;

le *coefficient de dissymétrie* : $\beta_1 = (e^{\sigma^2} - 1)^{\frac{1}{2}} (e^{\sigma^2} + 2)$;

le *coefficient d'aplatissement* : $\beta_2 = (e^{\sigma^2} - 1)(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6)$.

On remarque que les coefficients de dissymétrie et d'aplatissement de la loi lognormale sont strictement positifs (et d'autant plus proches de 0 que σ^2 l'est), et que le mode, la médiane et la moyenne se suivent dans l'ordre indiqué sur la figure 5.8 et, en outre, tendent à se confondre (et la courbe de la densité à devenir symétrique) à mesure que σ^2 se rapproche de 0. Ajoutons que bien que les valeurs prises par la variable lognormale soient les « perturbations aléatoires » d'une valeur « vraie » a, celle-ci *joue le rôle non pas de moyenne mais de médiane.*

6.1.7. Loi uniforme (rectangulaire). On dit qu'une variable aléatoire ξ est uniformément répartie sur un intervalle $[a, b]$ si sa densité de probabilité $f_\xi(x)$ est constante sur $[a, b]$ et nulle en dehors, c'est-à-dire que

$$f_\xi(x) = \begin{cases} \dfrac{1}{b-a} & \text{pour } x \in [a, b]; \\ 0 & \text{sinon.} \end{cases} \tag{6.13}$$

Cette loi s'appelle aussi *rectangulaire* car le graphique de la fonction $f_\xi(x)$ a la forme d'un rectangle (cf. fig. 6.2).

La fonction de répartition $F_\xi(x)$ de la loi uniforme est définie par les relations

$$F_\xi(x) = \begin{cases} 0 & \text{pour } x \leqslant a; \\ \dfrac{x-a}{b-a} & \text{pour } a < x \leqslant b; \\ 1 & \text{pour } x > b. \end{cases} \tag{6.13'}$$

Les variables aléatoires uniformes interviennent dans l'analyse des erreurs d'arrondi (ces erreurs sont en général uniformément réparties sur un intervalle compris entre -5 et $+5$ unités de la décimale arrondie); dans la description de la durée d'attente d'un service rendu dans une station fonctionnant avec une période rigou-

reuse T et une arrivée aléatoire des unités (exemple le temps d'attente de rames de métro se suivant à un intervalle de 2 mn et avec une arrivée aléatoire des passagers est une variable aléatoire uniformément répartie sur l'intervalle [0 mn, 2 mn]).

Signalons encore deux situations importantes faisant intervenir la loi uniforme. Premièrement, en analyse statistique des données, en théorie et en pratique on passe souvent de la variable aléatoire étudiée ξ de fonction de répartition $F(x)$ à la variable aléatoire $\eta = F(\xi)$ qui est uniformément répartie sur l'intervalle [0, 1] (cf. § 7.4). Ce procédé est payant dans la *modélisation statistique des observations* suivant une loi de probabilité donnée (cf. § 6.3), la construction de l'intervalle de confiance pour la fonction de répartition étudiée et dans un grand nombre d'autres problèmes de statistique mathématique. Deuxièmement, la loi uniforme est parfois utilisée à titre d'« approximation zéro » dans la description de la répartition de paramètres analysés dans le cadre de l'*approche bayesienne* en l'absence de toute information *a priori* sur cette répartition (cf. n° 8.6.6).

Les caractéristiques numériques de la loi uniforme sont :

la *moyenne*, la *médiane* : $\mathrm{E}\xi = x_{\text{méd}} = \dfrac{a+b}{2}$;

la *variance* : $\mathrm{V}\xi = \dfrac{(b-a)^2}{12}$

le *coefficient de dissymétrie* : $\beta_1 = 0$;

le *coefficient d'aplatissement* : $\beta_2 = -1,2$.

Signalons en conclusion une importante propriété de la somme de n variables aléatoires uniformes : la loi de cette somme tend très vite (à mesure que le nombre de termes croît) vers la loi normale. En particulier, si ξ_i sont des variables aléatoires indépendantes uniformément réparties sur l'intervalle [0, 1], la densité $f_{\eta_n}(x)$ de la variable aléatoire $\eta_n = \xi_1 + \ldots + \xi_n$ est de la forme

$$f_{\eta_n}(x) = \begin{cases} \dfrac{1}{(n-1)!}\, x^{n-1} & \text{pour } 0 \leqslant x \leqslant 1 ; \\[2ex] \dfrac{1}{(n-1)!}\,[x^{n-1} - C_n^1 (x-1)^{n-1}] & \text{pour } 1 < x \leqslant 2 ; \\[2ex] \dfrac{1}{(n-1)!}\,[x^{n-1} - C_n^1 (x-1)^{n-1} + C_n^2 (x-2)^{n-1}] & \\ & \text{pour } 2 < x \leqslant 3 ; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\[2ex] \dfrac{1}{(n-1)!}\,[x^{n-1} - C_n^1 (x-1)^{n-1} + \ldots + (-1)^{n-1} C_n^{n-1} (x-(n- & \\ & \text{pour } n-1 < x \leqslant n. \end{cases}$$

(le domaine des valeurs possibles de la variable aléatoire η_n est visiblement l'intervalle $[0, n]$). La figure 6.2 nous montre comment la forme de la densité $f_{\eta_n}(x)$ se modifie lorsque le nombre de termes n croît ($n = 1, 2, 3$). Cette propriété est notamment utilisée dans la modélisation statistique des observations suivant la loi normale.

6.1.8. Loi de Weibull et loi exponentielle. Considérons un mécanisme de formation de répartitions de variables aléatoires caractérisant, en particulier, la durée de vie d'un élément, d'un système

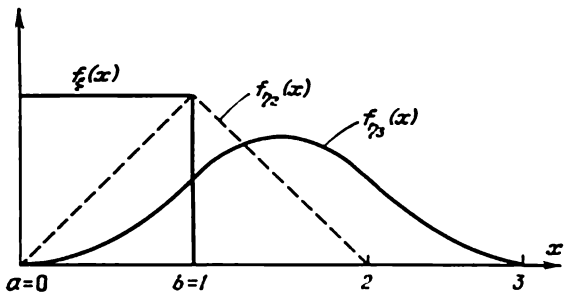


Fig. 6.2. Densités de probabilité ($f_\xi(x)$) d'une variable aléatoire uniforme, de la somme de deux ($f_{\eta_2}(x)$) et de trois ($f_{\eta_3}(x)$) variables aléatoires indépendantes uniformément réparties sur $[0, 1]$

complexe ou d'un individu (ces problèmes relèvent de la théorie de la fiabilité, de l'analyse des quotients de mortalité en démographie, etc.). Soit ξ la durée de vie de l'objet analysé (un système ou un individu) et soit $F(t) = P\{\xi < t\}$ sa fonction de répartition que nous supposerons continue et dérivable. Dans les problèmes de cette nature, une importante caractéristique est l'intensité des défaillances (le *quotient de mortalité*) $\lambda(t)$ des éléments d'âge t qui est définie par la relation

$$\lambda(t) = \frac{f_\xi(t)}{1 - F_\xi(t)} \approx \frac{n(t) - n(t + \Delta t)}{\Delta t \cdot n(t)}, \tag{6.14}$$

où $n(t)$ est le nombre d'objets ayant « vécu » jusqu'à l'âge t, Δt un intervalle de temps assez petit. Cela veut dire que statistiquement (empiriquement) le quotient de mortalité se définit comme le rapport du nombre spécifique (c'est-à-dire rapporté à l'unité de temps) des éléments « disparus » à l'âge t au nombre total $n(t)$ d'éléments ayant atteint cet âge. La véracité de l'égalité approchée (6.14) résulte de la relation

$$\frac{n(t) - n(t + \Delta t)}{\Delta t \cdot n(t)} = \frac{\dfrac{[n - n(t + \Delta t)] - [n - n(t)]}{\Delta t \cdot n}}{\dfrac{n - (n - n(t))}{n}} = \frac{\hat{f}_\xi^{(n)}(t)}{1 - \hat{F}_\xi^{(n)}(t)},$$

où n est le nombre total d'éléments ayant le même âge t, $n - n(t)$ le nombre de ceux qui n'ont pas atteint l'âge t (le rapport $(n - n(t))/n$ définit la fréquence relative de l'événement $\{\xi < t\}$).

La résolution de l'équation (6.14) par rapport à la fonction de répartition $F_\xi(t)$ nous donne

$$F_\xi(t) = 1 - e^{-\int_0^t \lambda(t)\,dt}. \tag{6.15}$$

Donc, la forme de la fonction de répartition $F_\xi(t)$ dépend entièrement de celle de $\lambda(t)$ (de la dépendance du quotient de mortalité par rapport au temps).

Les innombrables données empiriques (en démographie, analyse de la fiabilité des éléments et systèmes techniques, etc.) montrent

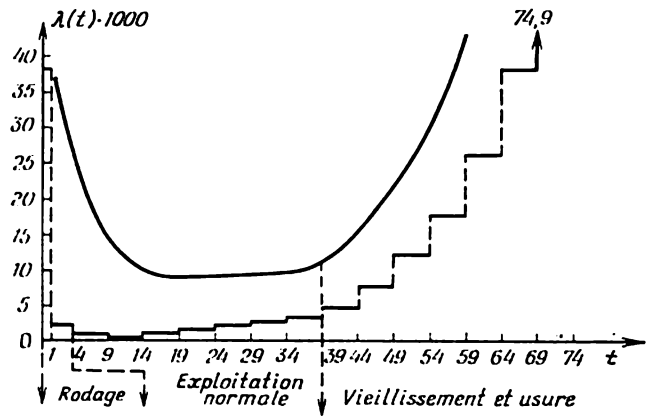


Fig. 6.3. Comportement typique de la courbe de mortalité (intensité des défaillances) et courbe réelle (en escalier) de variation du quotient de mortalité de la population masculine française en 1955

que dans une foule de cas la fonction $\lambda(t)$ est du type de la courbe représentée sur la figure 6.3. Sur ce graphique on voit que l'intervalle temporel peut être partagé en trois périodes. Sur la première, la fonction $\lambda(t)$ prend des valeurs élevées et a nettement tendance à décroître. Dans le langage technique ceci s'explique par la présence dans la population générale analysée d'éléments présentant des défauts manifestes ou latents (de montage, non-réalisation de certaines propriétés, etc.) qui précipitent l'usure de ces éléments. Cette période s'appelle *période de rodage*. La deuxième période est une période d'exploitation normale caractérisée par un niveau relativement faible et à peu près constant de « mortalité » des éléments. La mort (ou la « défaillance ») est subite (pannes, accidents, etc.)

pendant cette période et ne dépend pas de l'âge de l'élément. La troisième période, enfin, est la dernière période de vie (ou d'exploitation) de l'élément appelée encore *période de vieillissement et d'usure*. Les défaillances relèvent ici de phénomènes physiologiques ou physico-chimiques irréversibles qui conduisent à une détérioration de l'élément, à son « vieillissement ». Sous la courbe abstraite est construite une courbe en escalier réelle décrivant la variation du quotient de mortalité de la population masculine en France en 1955 *). Sur l'axe des ordonnées sont portés les nombres de décès pour mille personnes d'un âge donné. Conformément à cette courbe de mortalité, la période de « rodage » est comprise entre 0 et 4 ans, la période d'« exploitation normale » entre 4 et 39 ans, la période de « vieillissement », entre 39 ans et . . .

A chaque période est associée une fonction $\lambda(t)$, donc, une loi de probabilité du temps de vie ξ.

Considérons la classe des *fonctions puissances* $\lambda(t)$:

$$\lambda(t) = \lambda_0 \cdot \alpha t^{\alpha-1}, \tag{6.16}$$

où $\lambda_0 > 0$ et $\alpha > 0$ sont des paramètres numériques. Il est évident que les valeurs $\alpha < 1$, $a = 1$ et $\alpha > 1$ correspondent au comportement de la fonction d'intensité des défaillances pendant la période respectivement de rodage, d'exploitation normale et de vieillissement.

En portant (6.16) dans (6.15), on obtient la fonction de répartition $F_\xi(t)$ sous la forme suivante:

$$F_\xi(t) = 1 - e^{-\lambda_0 t^\alpha}, \quad t \geqslant 0. \tag{6.17}$$

La densité de probabilité est respectivement

$$f_\xi(t) = \lambda_0 \cdot \alpha \cdot t^{\alpha-1} \cdot e^{-\lambda_0 t^\alpha}, \quad t \geqslant 0. \tag{6.17'}$$

Ceci est la *loi de Weibull*. On aboutit à cette loi en étudiant la répartition des termes extrêmes des séries variationnelles avec des lois diverses (cf., par exemple, [65]).

Les principales caractéristiques numériques de la loi de Weibull sont:

la *moyenne*: $E\xi = \lambda_0^{-\frac{1}{\alpha}} \cdot \Gamma\left(1 + \frac{1}{\alpha}\right)$;

le *mode*: $x_{\text{mod}} = \begin{cases} 0 & \text{si } \alpha \leqslant 1; \\ \lambda_0^{-\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right)^{\frac{1}{\alpha}} & \text{si } \alpha \geqslant 1; \end{cases}$

*) Cf. R. Pressat. L'analyse démographique. Méthodes, résultats, applications. Presses Universitaires de France, 108 Boulevard Saint-Germain Paris, 1961.

la *variance*: $V\xi = \lambda_0^{-\frac{2}{\alpha}} \cdot \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$;

le *moment d'ordre k*: $m_k = E\xi^k = \lambda_0^{-\frac{k}{\alpha}} \cdot \Gamma\left(1+\frac{k}{\alpha}\right)$ où $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$ est la fonction gamma.

La *loi exponentielle* bien qu'étant un cas particulier (pour $\alpha = 1$) de la loi de Weibull n'en présente pas moins en soi un intérêt indiscutable. Nous avons vu qu'elle se prêtait bien à la description de la répartition de la durée de vie d'un élément fonctionnant en régime d'exploitation normale. La loi exponentielle (et elle seule) possède, en particulier, la remarquable propriété suivante: la probabilité de fonctionnement d'un élément sur un intervalle de temps $[t, t+\Delta]$ ne dépend pas du temps t antérieur mais uniquement de la longueur de l'intervalle Δ. La variable aléatoire exponentielle peut encore être interprétée comme une variable décrivant l'intervalle de temps séparant la réalisation successive d'un événement « poissonnien ». Le succès de la loi exponentielle tient non seulement à ses diverses interprétations mais aussi à l'extrême simplicité et commodité de ses propriétés. La fonction de répartition, la densité de probabilité et les principales caractéristiques numériques de cette loi sont:

$$F_\xi(x) = 1 - e^{-\lambda_0 x}, \quad x \geqslant 0;$$
$$f_\xi(x) = \lambda_0 e^{-\lambda_0 x}, \quad x \geqslant 0;$$

la *moyenne*: $E\xi = \frac{1}{\lambda_0}$;

le *mode*: $x_{\text{mod}} = 0$;

la *médiane*: $x_{\text{méd}} = \frac{1}{\lambda_0} \cdot \ln 2$;

la *variance*: $V\xi = \frac{1}{\lambda_0^2}$;

le *coefficient de dissymétrie*: $\beta_1 = 2$;

le *coefficient d'aplatissement*: $\beta_2 = 6$.

Loi exponentielle bilatérale (loi de Laplace). La densité de probabilité de cette loi qui est unimodale symétrique et avec un maximum en « pointe » en $x = 0$ est souvent utilisée pour décrire la répartition des composantes aléatoires résiduelles ε (des « erreurs ») dans des modèles de type (3.5) et (3.9). Le graphique de cette densité s'obtient en « collant » la courbe de la loi exponentielle à sa symétrique par rapport à un axe vertical (compte tenu de la normalisation requise), de sorte que l'équation de la courbe de la densité est

$$f(x) = \frac{1}{2}\lambda \cdot e^{-\lambda \cdot |x|} \quad (-\infty < x < \infty).$$

Il est immédiat de calculer les principales caractéristiques numériques de cette loi :

la *moyenne* : $E\xi = 0$;
le *mode* : $x_{mod} = 0$;
la *médiane* : $x_{méd} = 0$;
la *variance* : $V\xi = 2/\lambda^2$;
le *coefficient de dissymétrie* : $\beta_1 = 0$;
le *coefficient d'aplatissement* : $\beta_2 = 3$.

6.1.9. Loi de Pareto. Les lois « tronquées » se rencontrent assez souvent dans les divers problèmes de statistique appliquée. Ces lois décrivent le comportement de populations générales « tronquées », c'est-à-dire dont on a retiré tous les éléments en deça ou au-delà d'un certain indice c_0. Le fisc, par exemple, impose les revenus annuels supérieurs à un certain taux x_0. Cette situation se décrit approximativement par la loi de Pareto dont la fonction de répartition et la densité sont :

$$F_\xi(x) = P\{\xi < x\} = 1 - \left(\frac{c_0}{x}\right)^\alpha,$$

$$f_\xi(x) = \frac{\alpha}{c_0}\left(\frac{c_0}{x}\right)^{\alpha+1}.$$

Dans ces formules, $\alpha > 0$ et $x > c_0$, c'est-à-dire que le domaine des valeurs possibles de la variable aléatoire ξ est la demi-droite $]c_0, \infty[$. La densité est une fonction monotone décroissante dont la courbe est issue du point $(c_0, \alpha/c_0)$.

Les principales caractéristiques numériques de cette loi n'existe que si le paramètre α satisfait à certaines conditions :

la *moyenne* : $E\xi = \frac{\alpha}{\alpha-1} c_0$ (existe pour $\alpha > 1$) ;

le *mode* : $x_{mod} = c_0$;

la *médiane* : $x_{méd} = 2^{\frac{1}{\alpha}} \cdot c_0$;

la *variance* : $V\xi = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} c_0^2$ (existe pour $\alpha > 2$) ;

le *moment d'ordre k* : $E\xi^k = \frac{\alpha}{\alpha-k} c_0^k$ (existe pour $\alpha > k$).

Pour de plus amples renseignements sur la loi de Pareto, voir : Hagstroem. La loi de Pareto et la réassurance. — Skandinavisk Akturiet (1925), 65 ; Inkomstutjämningen i Sverige. — Skandinavsk Bankens Kvart, skr., April (1944).

6.1.10. Loi de Cauchy. Cette loi est assez spécifique dans la mesure où aucun de ses moments d'ordre positif (y compris la moyenne) n'existe. La loi de Cauchy est unimodale, symétrique par rapport à son mode (qui, par conséquent, est aussi la médiane) et a pour

densité de probabilité

$$f(x)=\frac{1}{\pi}\cdot\frac{c}{c^2+(x-a)^2}, \quad -\infty<x<\infty$$

où $c>0$ est le paramètre d'échelle, a le paramètre du centre de groupement qui définit en même temps le mode et la médiane.

La fonction de répartition est définie respectivement par la relation

$$F(x)=\frac{1}{2}+\frac{1}{\pi}\operatorname{Arctg}\frac{x-a}{c}.$$

Signalons deux importantes propriétés (d'« autoreproduction ») de la loi de Cauchy (cf., par exemple, [40]):

1. Si une variable aléatoire ξ suit la loi de Cauchy avec les paramètres c et a, alors toute fonction linéaire $b_0+b_1\xi$ suit la même loi avec les paramètres $c'=|b_1|\cdot c$ et $a'=b_1\cdot a+b_0$;

2. Si des variables aléatoires $\xi_1, \xi_2, \ldots, \xi_n$ sont indépendantes et obéissent à la même loi de Cauchy, leur moyenne arithmétique $\bar{\xi}=(\xi_1+\ldots+\xi_n)/n$ suivra aussi cette loi.

6.1.11. Quelques combinaisons des principales lois de probabilité modèles utilisées en statistique appliquée. On peut obtenir des lois de probabilité en combinant les lois modèles. Certaines des plus « commodes » sur le plan de la méthodologie et des applications sont décrites plus bas.

Un *modèle proche de la loi normale mais tenant compte de la non-nullité des coefficients de dissymétrie* (β_1) *et d'aplatissement* (β_2) (cf. n° 5.6.6) *peut être défini par la densité*

$$f(x)=\varphi(x)-\frac{\beta_1}{3!}\varphi^{(3)}(x)+\frac{\beta_2}{4!}\varphi^{(4)}(x),$$

où $\varphi(x)$ est la densité de la loi normale et $\varphi^{(k)}(x)$ ses dérivées. Ce modèle est engendré par des développements asymptotiques dans le théorème limite central (cf. [40]) qui montrent comment la loi d'une somme de variables aléatoires indépendantes se rapproche de la loi normale. On voit sur la figure 6.4 comment la non-nullité des coefficients de dissymétrie et d'aplatissement influe sur la densité $f(x)$. A noter que l'expression de la fonction $f(x)$ a été établie sur des considérations *asymptotiques*, donc, d'une façon générale, elle ne définit pas la densité de probabilité pour toutes les valeurs de β_1 et β_2.

Pour être fixés sur les valeurs que peuvent prendre les coefficients β_1 et β_2 dans les applications, signalons, suivant [117], que $-0{,}9\leqslant\beta_1\leqslant 0{,}9$ et $-0{,}6\leqslant\beta_2\leqslant 1{,}8$ pour de nombreuses lois utilisées en technique, en biologie, en chimie et en météorologie. En démographie, par exemple, ces coefficients prennent des valeurs très grandes; ainsi, ils varient dans l'intervalle (2,0; 6,3) dans la ré-

partition de l'âge des femmes qui se sont mariées entre 1907 et 1914 en Australie, et dans l'intervalle (2,0; 5,3), pour les hommes.

Le *modèle d'un mélange de lois de type donné* $f_j(x;\ \theta_j)$ est décrit par la formule

$$f(x) = \sum_{j=1}^{k} \pi_j \cdot f_j(x;\ \theta_j),$$

dans laquelle $f_j(x;\ \theta_j)$ et $f(x)$ sont les densités (dans le cas continu) ou les polygones de fréquence (dans le cas discret) respectivement de

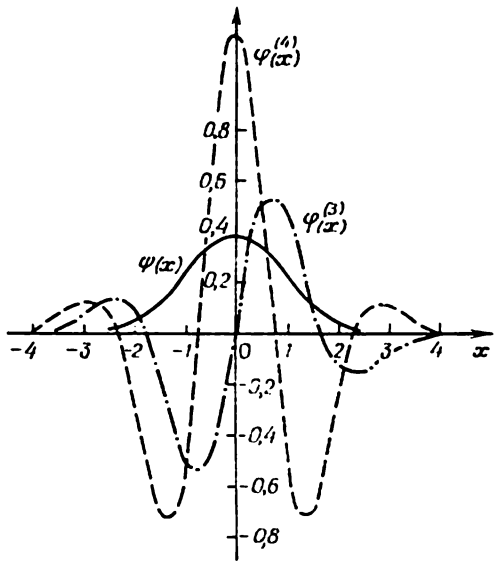


Fig. 6.4. Densité de la loi normale $\varphi(x)$ et ses dérivées $\varphi^{(3)}(x)$ et $\varphi^{(1)}(x)$

la j-ième composante du mélange et de la loi de probabilité résultante, π_j, la probabilité *a priori* d'apparition dans un échantillon aléatoire d'une observation de loi $f_j(x;\ \theta_j)$ (c'est-à-dire le poids spécifique de telles observations dans la population générale), k le nombre de composantes du mélange. L'analyste est confronté à de telles lois de probabilité lorsque, par exemple, il est amené à analyser une population générale composée de plusieurs sous-populations qui tout en étant homogènes dans un certain sens (ce qui peut s'exprimer, par exemple, par l'*unimodalité* de la loi de probabilité $f_j(x;\ \theta_j)$) sont fondamentalement différentes l'une de l'autre (par exemple, par la valeur du paramètre θ_j). Ceci étant, le paramètre θ_j peut définir aussi bien le centre de groupement des observations correspondantes (auquel cas il s'interprète comme un paramètre de localisation) que

le degré de leur dispersion aléatoire (il est alors interprété comme un paramètre d'échelle).

On peut trouver de plus amples informations sur les mélanges des lois dans [9]. Dans [43] et [137] sont décrits des exemples de mécanismes naturels réels empruntés à l'économie et à la nature, qui amènent à considérer un mélange de lois.

Rappelons quelques cas particuliers de modèles de mélanges utilisés par de nombreux chercheurs pour établir des inférences statistiques stables.

Le *modèle de Tuckey de la loi normale « pollué »* est appliqué, par exemple, à l'étude de l'influence des ailes allourdies de la répartition sur les propriétés des estimations de la moyenne inconnue (cf. n° 8.6.4). Ceci étant, on part du fait que les observations sont « extraites » d'une population générale définie par une densité de probabilité de la forme

$$f(x) = (1 - \varepsilon)\,\varphi(x;\ a;\ \sigma_0^2) + \varepsilon\cdot\varphi(x;\ a,\ \sigma_1^2),$$

où $\varphi(x;\ a,\ \sigma^2)$ est la densité de la loi normale; ε le taux (en général, relativement peu élevé) des observations « polluantes » et où les variances des deux composantes sont reliées par l'inégalité $\sigma_0^2 < \sigma_1^2$.

Modèle de pollution de Chouryguine. Les pollutions rencontrées dans la pratique sont souvent dissymétriques. Pour traduire ce fait, on peut introduire dans le modèle du mélange de lois un paramètre supplémentaire a exprimant l'écart de la pollution par rapport à la loi principale de densité $\psi(x,\ \theta,\ \sigma)$. Il faut alors envisager un modèle de mélange de la forme

$$f(x) = (1 - \varepsilon)\,\psi(x,\ \theta,\ \sigma) + \varepsilon h(x - \theta - a),$$

où θ et σ sont respectivement les paramètres du centre de groupement (de localisation) et d'échelle, h la densité d'une loi symétrique. Pour lever l'arbitraire qui préside au choix de h et présenter les résultats obtenus pour diverses valeurs de a sous une forme moins volumineuse, on a proposé (cf. [34]) de considérer le *schéma des séries* d'épreuves suivantes: dans chaque série on prélève un échantillon ordinaire suivant un mélange de lois, ceci étant, on admet par souci de simplicité que la pollution est toujours concentrée *en un point* $(\theta + a)$, mais au passage d'une série à une autre, le paramètre a est choisi de façon aléatoire et suit une loi normale de moyenne nulle et d'écart-type $k\sigma$. Le modèle de Chouryguine s'est avéré plus commode pour l'analyse.

6.2. Lois de probabilité utilisées dans les calculs statistiques

6.2.1. Loi du χ^2. On décrit plus bas cinq lois de probabilité qui se prêtent particulièrement bien à la construction de divers tests statistiques et à l'estimation par intervalle des paramètres:

la loi du « khi deux »; la loi t de Student; la loi F de Fisher dite encore loi du rapport dispersionnel; la loi B (« bêta »); la loi Γ (« gamma »). L'explication du mécanisme d'action de ces lois peut être liée à la « statistique de la loi normale », c'est-à-dire à l'étude des répartitions de certaines fonctions de variables aléatoires indépendantes suivant la loi normale réduite.

L'astronome F. Helmert a appliqué la théorie gaussienne des erreurs pour étudier la somme des carrés de variables aléatoires normales. Cette étude l'a conduit à une fonction de répartition $F_{\chi^2(m)}(x)$ que plus tard K. Pearson a appelée fonction de répartition de χ^2 *) et qui est définie par

$$F_{\chi^2(m)}(x) = \begin{cases} 0 \text{ pour } x < 0, \\ P\{\chi^2(m) < x\} = \dfrac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} \cdot \displaystyle\int_0^x t^{\frac{m}{2}-1} e^{-\frac{t}{2}} dt \\ \qquad\qquad \text{pour } x \geqslant 0, \end{cases} \tag{6.18}$$

où m est un paramètre >0 appelé *nombre de degrés de liberté*, $\Gamma(y)$ la valeur de la fonction gamma au point y.

La densité de probabilité correspondante est définie par

$$f_{\chi^2(m)}(x) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} \cdot x^{\frac{m}{2}-1} e^{-\frac{x}{2}}, \quad x > 0. \tag{6.18'}$$

Pour $m \leqslant 2$ la densité décroît constamment (si $x > 0$) et pour $m > 2$ présente un seul maximum au point $x_{\text{mod}} = m - 2$.

La loi du χ^2 est apparue pour la première fois dans l'étude de la répartition d'une suite de variables aléatoires indépendantes $\xi_1, \xi_2, \ldots, \xi_m$ suivant la loi normale réduite. Il s'est avéré que la variable aléatoire $\chi^2(m) = \xi_1^2 + \ldots + \xi_m^2$ suit la loi du χ^2 à m degrés de liberté. Ceci entraîne l'important résultat suivant: si $s^2(n) = \sum_{i=1}^{m} (x_i - \bar{x}(n))^2/n$ est la variance empirique construite avec les observations indépendantes $x_1, x_2, \ldots, x_n$ d'une variable aléatoire normale de paramètres (a, σ^2), alors la variable aléatoire $ns^2(n)/\sigma^2$ est répartie selon la loi du χ^2 à $n - 1$ degrés de liberté, soit

$$\frac{ns^2(n)}{\sigma^2} = \chi^2(n-1). \tag{6.19}$$

*) Helmert F.R. *Über die Wahrscheinglichkeit von Potenzsummen der Beobachtungsfehler* etc. Z.F. Math. und Phys., 21 (1876); Pearson K. *On the criterion that a given system of deviation from the probable in the case of correlated system of variables...*, « Phil Mag. », V. 50 (1900), 157.

Citons encore deux importants résultats acquis en appliquant la loi du χ^2 au traitement statistique des données.

Supposons que la variable aléatoire ξ admet une fonction de répartition $F_\xi(x;\ \theta_1, \ldots, \theta_s)$ dépendant *suffisamment différentiablement* *) de s paramètres inconnus $\theta_1, \ldots, \theta_s$. Supposons qu'on ait réussi à construire à l'aide de l'échantillon $x_1, x_2, \ldots, x_n$ des estimations *assez bonnes* (efficaces ou asymptotiquement efficaces, cf. § 8.3) $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$ pour les valeurs inconnues des paramètres $\theta_1, \theta_2, \ldots, \theta_s$ (dans la loi normale les paramètres θ_1 et θ_2 ne sont autres respectivement que la moyenne m et la variance σ^2, et leurs estimations, des fonctions des observations $\bar{x}(n)$ et $s^2(n)$). Définissons par ailleurs les probabilités:

$$p_i = \begin{cases} P\{\xi = x_i^0\} = F_\xi(x_{i+1}^0;\ \hat{\theta}_1, \ldots, \hat{\theta}_s) - \\ \qquad\qquad - F_\xi(x_i^0;\ \hat{\theta}_1, \ldots, \hat{\theta}_s) \quad \text{si } \xi \text{ est discret;} \\ F_\xi(\tilde{x}_{i+1}^0;\ \hat{\theta}_1, \ldots, \hat{\theta}_s) - \\ \qquad\qquad - F_\xi(\tilde{x}_i^0;\ \hat{\theta}_1, \ldots, \hat{\theta}_s) \quad \text{si } \xi \text{ est continu,} \end{cases}$$

où x_i^0 et $\tilde{x}_i^0$ sont respectivement la i-ième valeur possible de ξ et l'extrémité gauche du i-ième intervalle de groupement ($i = 1, 2, \ldots, k$); k le nombre total de valeurs possibles ou d'intervalles de groupement, les valeurs de F_ξ aux extrémités étant égales à 0 (en x_1^0 ou $\tilde{x}_1^0$) et à 1 (en x_{k+1}^0 ou $\tilde{x}_{k+1}^0$). Si ν_i ($i = 1, 2, \ldots, k$) est le nombre d'observations égales à x_i^0 (ou tombant dans le i-ième intervalle de groupement), la loi de la mesure intégrale de l'écart entre les fréquences relatives empiriques $\frac{\nu_i}{n}$ et les probabilités correspondantes p_i

$$\chi^2(k-s-1) = n \sum_{i=1}^{k} \frac{\left(p_i - \frac{\nu_i}{n}\right)^2}{p_i} = \sum_{i=1}^{k} \frac{(np_i - \nu_i)^2}{np_i} \tag{6.20}$$

tend vers la loi du χ^2 à $k - s - 1$ degrés de liberté lorsque $n \to \infty$. Ce résultat est appliqué pour vérifier les hypothèses statistiques concernant la forme de la loi analysée (cf. § 11.1).

Si dans les conditions et les notations du résultat précédent, on ajoute un deuxième échantillon parent $x_1', x_2', \ldots, x_{n'}'$ (les fré-

*) Pour que le résultat ultérieur regroupant les théorèmes de K. Pearson (1900) et R. Fisher (1922) soit valable, il suffit d'exiger l'existence des dérivées partielles continues $\frac{dF_\xi}{d\theta_j}$ ($j = 1, 2, \ldots, s$).

quences relatives respectives étant ν_i'/n'), la quantité

$$\chi^2(k-1) = n \cdot n' \sum_{i=1}^{k} \frac{1}{\nu_i + \nu_i'} \left(\frac{\nu_i}{n} - \frac{\nu_i'}{n'}\right)^2 \tag{6.21}$$

suivra une loi qui tendra vers la loi du χ^2 à $k-1$ degrés de liberté lorsque $\min(n, n') \to \infty$. Ce résultat est utilisé pour vérifier l'hypothèse d'homogénéité de deux échantillons distincts et peut être généralisé à plus de deux échantillons (cf. § 11.2).

Les principales caractéristiques numériques de la loi du $\chi^2(m)$ sont :

la *moyenne* : $\mathrm{E}\chi^2(m) = m$;
le *mode* : $x_{\mathrm{mod}} = m - 2$ $(m \geqslant 2)$;
la *variance* : $\mathrm{V}\chi^2(m) = 2m$;
le *coefficient de dissymétrie* $\beta_1 = \dfrac{2^{3/2}}{\sqrt{m}}$;
le *coefficient d'aplatissement* : $\beta_2 = \dfrac{12}{m}$.

6.2.2. La loi t de Student. En analysant les écarts aléatoires de la moyenne empirique $\bar{x}(n)$ par rapport à la moyenne théorique $m_1 = \mathrm{E}\xi$ de ξ, le statisticien anglais V. Gosset (qui écrivait sous le nom de plume de Student) a obtenu le résultat suivant en 1908 *). Soient $\xi_0, \xi_1, \ldots, \xi_m$ des variables aléatoires indépendantes normales de paramètres $(0, \sigma^2)$. La densité de probabilité de la variable aléatoire

$$t(m) = \frac{\xi_0}{\left(\frac{1}{m}\sum_{i=1}^{m} \xi_i^2\right)^{\frac{1}{2}}}$$

est

$$f_t(x) = \frac{1}{\sqrt{\pi m}} \cdot \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \cdot \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \quad (-\infty < x < +\infty). \tag{6.22}$$

La loi (6.22) s'appelle *loi de Student à m degrés de liberté* (ou loi $t(m)$). On voit que la densité (6.22) ne dépend pas de la variance σ^2 des variables ξ_i et, de plus, est unimodale et symétrique par rapport au point $x = 0$.

Citons quelques résultats utilisés pour le traitement statistique des données empiriques extraites d'une population générale normale.

1. Si $x_1, x_2, \ldots, x_n$ est un échantillon d'une population générale normale de paramètres (a, σ^2), et $\bar{x}(n)$ et $s^2(n)$ sont respectivement

*) Student. *The probable error of a mean.* — Biometrika, B. 6 (1908), 1.

la moyenne et la variance empiriques, construites sur les observations de cet échantillon, alors la variable aléatoire

$$t(n-1) = \frac{(\bar{x}(n)-a)\sqrt{n-1}}{s} \tag{6.23}$$

est répartie suivant la loi de Student à $n-1$ degrés de liberté. Ce résultat est utilisé dans l'estimation des intervalles de variation du paramètre inconnu $a = \mathrm{E}\xi$, ainsi que pour vérifier l'hypothèse statistique concernant l'appartenance de l'échantillon considéré à une population normale de moyenne donnée $\mathrm{E}\xi = a$ (cf. chap. 8 et 9).

2. Si, en se plaçant dans les conditions et les notations de l'exemple précédent, on considère un deuxième échantillon parent $x'_1, x'_2, \ldots, x'_{n'}$ (la moyenne et la variance empiriques de cet échantillon étant respectivement $\bar{x}'(n')$ et $s'^2(n')$), alors la mesure normalisée de l'écart des deux moyennes empiriques

$$t(n+n'-2) = \frac{(\bar{x}(n)-\bar{x}'(n'))\sqrt{\dfrac{nn'}{n+n'}}}{\bar{s}(n, n')} \tag{6.24}$$

suit la loi de Student à $n+n'-2$ degrés de liberté (l'écart-type commun $\bar{s}(n, n')$ de la formule (6.24) est défini par la relation

$$\bar{s}^2(n, n') = \frac{n}{n+n'-2}s^2(n) + \frac{n'}{n+n'-2}s'^2(n')).$$

Ce résultat est utilisé pour vérifier l'homogénéité des moyennes empiriques construites sur deux échantillons distincts extraits de populations normales (cf. n° 11.2.5).

Les principales caractéristiques numériques de la loi $t(m)$ sont :

la *moyenne*, le *mode*, la *médiane* : $\mathrm{E}t(m) = x_{\text{mod}} = x_{\text{méd}} = 0$;

la *variance* : $\mathrm{V}t(m) = \dfrac{m}{m-2}$ (n'existe que pour $m > 2$) ;

le *coefficient de dissymétrie* : $\beta_1 = 0$;

le *coefficient d'aplatissement* : $\beta_2 = \dfrac{6}{m-4}$ (n'existe que pour $m > 4$).

6.2.3. Loi F (loi du rapport dispersionnel). L'analyse du comportement du rapport de deux variances empiriques $s^2(n)$ et $s'^2(n')$ construites sur les observations de deux échantillons $x_1, x_2, \ldots, x_n$ et $x'_1, x'_2, \ldots, x'_{n'}$ extraits de la *même population normale*, a conduit en 1924 le statisticien anglais R. Fisher à une loi *) qui a reçu le nom de *loi F* et qui dans le cas général peut être définie comme suit.

*) Fisher R. *On a distribution yielding the error functions of several well-known statistics.* — Proc. Intern. Math. Congr. Toronto, 1924, 805.

Considérons $m_1 + m_2$ variables indépendantes $\xi_1, \ldots, \xi_{m_1}$, $\eta_1, \ldots, \eta_{m_2}$ suivant une loi normale de paramètres $(0, \sigma^2)$ et posons

$$F(m_1, m_2) = \frac{\frac{1}{m_1}\sum_{i=1}^{m_1} \xi_i^2}{\frac{1}{m_2}\sum_{j=1}^{m_2} \eta_j^2}$$

Il est évident (cf. n° 6.2.1) que cette variable aléatoire peut également être définie comme le rapport de deux variables $\chi^2(m_1)$ et $\chi^2(m_2)$ indépendantes, convenablement normées, soit

$$F(m_1, m_2) = \frac{\frac{1}{m_1}\chi^2(m_1)}{\frac{1}{m_2}\chi^2(m_2)}.$$

On démontre que la densité de probabilité de la variable aléatoire $F(m_1, m_2)$ est

$$f_{F(m_1, m_2)}(x) = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)\cdot m_1^{\frac{m_1}{2}}\cdot m_2^{\frac{m_2}{2}}}{\Gamma\left(\frac{m_1}{2}\right)\cdot\Gamma\left(\frac{m_2}{2}\right)} \times$$

$$\times \frac{x^{\frac{m_1}{2}-1}}{(m_1 x + m_2)^{\frac{m_1+m_2}{2}}} \quad (0 \leqslant x < \infty), \qquad (6.25)$$

où comme toujours $\Gamma(y)$ est la valeur de la fonction gamma au point y. La loi définie par (6.25) s'appelle loi F à (m_1, m_2) degrés de liberté.

La loi F est utilisée de la manière suivante pour le traitement statistique des données.

1. Si $\bar{s}^2(n) = \sum_{i=1}^{n} (x_i - \bar{x}(n))^2/(n-1)$ et $\bar{s}'^2(n') = \sum_{i=1}^{n} (x_i - \bar{x}'(n'))^2$ sont les variances empiriques construites sur des échantillons indépendants $x_1, \ldots, x_n$ et $x'_1, \ldots, x'_{n'}$ extraits d'une même population normale, alors le rapport $\bar{s}^2(n)/\bar{s}'^2(n')$ obéit à une loi F à $(n-1, n'-1)$ degrés de liberté.

2. Soient $X_1, X_2, \ldots, X_n$ un échantillon d'une population normale p-dimensionnelle de vecteur des moyennes $\mathbf{E}X = M$ et de matrice des covariances $\mathbf{E}\{(X-m)(X-m)'\} = \Sigma$ et soient $\bar{X}(n)$ et $\mathbf{S}(n)$ le vecteur des moyennes empiriques et la matrice des covariances empiriques construits à l'aide de cet échantillon (cf.

n° 5.6.7). Alors la variable aléatoire

$$\frac{n}{n-1}\cdot\frac{n-p}{p}\cdot(\overline{X}-M)'\,S^{-1}(\overline{X}-M) \tag{6.26}$$

suit une loi F à $(p.\ n-p)$ degrés de liberté. Dans l'expression (6.26) le facteur

$$T^2 = n\,(\overline{X}\,(n)-M)'\,S^{-1}\,(\overline{X}\,(n)-M) \tag{6.26'}$$

est la généralisation de la t-statistique de Student (6.23) utilisée pour vérifier l'hypothèse concernant la valeur de la moyenne et peut être interprété comme la caractéristique de l'écart géométrique (au sens d'une métrique de Mahalanobis, cf. [16]) entre la moyenne empirique $\overline{X}\,(n)$ et la moyenne théorique M.

3. Considérons maintenant un deuxième échantillon $X'_1, \ldots, X'_{n'}$ de la *même* population normale p-dimensionnelle et les caractéristiques empiriques correspondantes: le vecteur des moyennes $\overline{X}'\,(n')$ et la matrice des covariances $S'(n')$. Introduisons la matrice des covariances empirique $\overline{S}\,(n,\ n')$ construite sur les deux échantillons:

$$\overline{S}\,(n,\ n') = \frac{n}{n+n'-2}\cdot S\,(n) + \frac{n'}{n+n'-2}\,S'\,(n').$$

L'*analogue multidimensionnel* de la t-statistique (6.24), utilisée pour tester l'homogénéité de deux moyennes empiriques, est la quantité

$$\widetilde{T}^2 = \frac{nn'}{n+n'}\,(\overline{X}\,(n)-\overline{X}'\,(n'))'\,\overline{S}^{-1}\,(n,\ n')\,((\overline{X}\,(n)-\overline{X}'\,(n')).$$

de plus, la variable aléatoire

$$\frac{n+n'-p-1}{(n+n'-2)\cdot p}\cdot\widetilde{T}^2 \tag{6.27}$$

est répartie suivant une loi F à $(p,\ n+n'-p-1)$ degrés de liberté.

Les principales caractéristiques numériques de la loi $F\,(m_1,\ m_2)$ sont:

la *moyenne*: $\mathbf{E}F\,(m_1,\ m_2) = \dfrac{m_2}{m_2-2}$ (si seulement $m_2 > 2$);

le *mode*: $x_{\text{mod}} = \dfrac{(m_1-2)\cdot m_2}{m_1\,(m_2+2)}$ (pour $m_1 > 1$);

la *variance*: $\mathbf{V}F\,(m_1,\ m_2) = \dfrac{2m_2^2\,(m_1+m_2-2)}{m_1\,(m_2-2)^2\,(m_2-4)}$ (pour $m_2 > 4$);

le *coefficient de dissymétrie*: $\beta_1 = \dfrac{(2m_1+m_2-2)\sqrt{8\,(m_2-4)}}{(m_2-6)\sqrt{m_1+m_2-2}}$ (pour $m_2 > 6$).

De là il s'ensuit immédiatement que pour $m_1,\ m_2 > 2$, la loi F admet toujours un mode <1 et une moyenne >1. Cela signifie, en

particulier, que cette loi admet un *coefficient de dissymétrie strictement positif* non seulement pour $m_2 > 6$ (ce qui résulte de la forme de β_1) mais aussi pour m_1, $m_2 > 2$.

Fisher a établi cette loi non pas pour la variable aléatoire $F(m_1, m_2)$, mais pour son logarithme népérien (divisé par deux), c'est-à-dire pour la variable aléatoire $\zeta = \frac{1}{2} \ln F(m_1, m_2)$. La loi de cette variable est souvent appelée *loi z de Fisher*. Mais les statisticiens préfèrent utiliser la loi F pour ses propriétés plus simples.

6.2.4. Remarque sur les lois non centrées du χ^2, F et t. Dans ce numéro, on introduit les lois appliquées à l'étude de la *puissance* des tests statistiques de modèles linéaires.

On dira que x suit la loi $N(\mu, \sigma^2)$ si x est une variable aléatoire normale de moyenne μ et de variance σ^2.

On appelle *loi du χ^2 non centrée à m degrés de liberté* la loi de la variable aléatoire $u = \sum_{i=1}^{m} x_i^2$, où $x_1, \ldots, x_m$ sont des variables aléatoires indépendantes suivant la loi normale $N(\mu_i, 1)$. La quantité $\delta = (\sum_{i=1}^{m} \mu_i^2)^{1/2}$ s'appelle *paramètre de décentrage* de cette loi. Cette loi est généralement désignée par $\chi^2(m, \delta)$, notation qui est correcte, puisque la loi de u ne dépend de $\mu_1, \ldots, \mu_m$ que par l'intermédiaire de δ. Lorsque $\delta = 0$, on dit que u suit une loi du χ^2 centrée qui est notée $\chi^2(m)$.

La *moyenne* et la *variance* de $\chi^2(m, \delta)$ sont :

$$\mathrm{E}(\chi^2(m, \delta)) = m + \delta^2, \quad \mathrm{V}\chi^2(m, \delta) = 2m + 4\delta^2.$$

Addition de variables aléatoires suivant une loi du χ^2 non centrée. Si $\chi^2(m_1, \delta_1)$ et $\chi^2(m_2, \delta_2)$ sont indépendantes, leur somme suit aussi une loi du $\chi^2(m, \delta)$, où $m = m_1 + m_2$ et $\delta = (\delta_1^2 + \delta_2^2)^{1/2}$.

Si les variables aléatoires $\chi^2(m_1, \delta)$ et $\chi^2(m_2)$ sont indépendantes, la loi du rapport $F(m_1, m_2; \delta) = \frac{\chi^2(m_1\delta)/m_1}{\chi^2(m_2)/m_2}$ s'appelle *loi F non centrée à (m_1, m_2) degrés de liberté et de paramètre de décentrage* δ.

Si x suit la loi normale $N(\mu, 1)$ et si la loi du $\chi^2(m)$ est indépendante de x, la loi du rapport $t(m, \delta) = \frac{x}{\sqrt{\chi^2(m)/m}}$ s'appelle *loi t de Student non centrée à m degrés de liberté et de paramètre de décentrage* $\delta = |\mu|$. Il est évident que $t^2(m, \delta) = F(1, m; \delta)$.

6.2.5. Loi gamma. Les deux lois qui seront décrites ultérieurement représentent des familles de lois à deux paramètres assez vastes et souples qui englobent diverses combinaisons des lois des variables déjà étudiées. Les lois gamma et bêta *sont importantes* surtout *par leurs applications dans les calculs* : en effet, elles peuvent servir à

calculer les lois du χ^2, t, F, etc. (cf., par exemple, [27]). De plus, la loi gamma est parfois utilisée en modélisation : elle décrit, par exemple, la répartition des revenus et des épargnes de la population dans des conditions spéciales (cf. [3], [4]).

La loi gamma à deux paramètres de la variable aléatoire $\gamma(a, b)$ est décrite par la densité de probabilité :

$$f_{\gamma(a,\, b)}(x) = \begin{cases} \dfrac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & \text{pour } x \in [0, \infty[\,; \\ 0 & \text{pour } x < 0, \end{cases} \tag{6.28}$$

où $\Gamma(a)$ est la fonction gamma, $a > 0$ le paramètre de « forme » et $b > 0$ le paramètre d'échelle. Il est immédiat qu'une variable aléatoire suivant une loi gamma de paramètres a et b (nous désignerons cette variable par $\gamma(a, b)$) est reliée à la variable aléatoire $\gamma(a, 1)$ par la relation élémentaire (cf. aussi § 7.4) :

$$b\gamma(a, b) = \gamma(a, 1).$$

Signalons quelques propriétés utiles de la loi gamma.

1. Un coup d'œil sur la formule (6.18′) de la densité de probabilité de la loi du $\chi^2(m)$ nous montre qu'elle est un cas particulier de la loi gamma : il suffit de poser $a = m/2$ et $b = {}^1/_2$ dans (6.28).

2. La somme d'un nombre quelconque de variables aléatoires $\gamma_1(a_1, b) + \gamma_2(a_2, b) + \ldots + \gamma_n(a_n, b)$ est aussi une variable aléatoire $\gamma(a_1 + \ldots + a_n, b)$.

Les principales caractéristiques numériques de $\gamma(a, b)$ sont :

la *moyenne* : $\mathrm{E}\gamma(a, b) = \dfrac{a}{b}$;

le *mode* : $x_{\mathrm{mod}} = \dfrac{a-1}{b}$ (pour $a \geqslant 1$) ;

la *variance* $\mathrm{V}\gamma(a, b) = \dfrac{a}{b^2}$;

le *coefficient de dissymétrie* : $\beta_1 = \dfrac{2}{\sqrt{a}}$;

le *coefficient d'aplatissement* : $\beta_2 = \dfrac{6}{a}$.

6.2.6. Loi bêta. On a signalé plus haut que la loi bêta à deux paramètres se distinguait par sa grande souplesse et son universalité : elle peut servir notamment à calculer les lois t^2, F, binomiale, binomiale négative, etc. Elle est utilisée aussi pour décrire certaines répartitions concentrées sur l'intervalle [0, 1] (par exemple, la répartition des probabilités subjectives obtenues par un sondage d'experts, cf. n° 4.1.3). Désignons par $\beta(a_1, a_2)$ une variable aléatoire suivant une loi bêta de paramètres a_1 et a_2 ($0 < a_1 < \infty$, $0 < a_2 <$

$< \infty$). Cette loi a pour densité

$$f_{\beta(a_1, a_2)}(x) = \begin{cases} \dfrac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} \cdot x^{a_1-1}(1-x)^{a_2-1} & \text{pour } x \in [0, 1]; \\ 0 & \text{pour tous les autres } x. \end{cases} \tag{6.29}$$

Signalons quelques propriétés utiles de la loi bêta.

1. Si $\gamma(a_1, b)$ et $\gamma(a_2, b)$ sont deux variables aléatoires indépendantes, le rapport $\gamma(a_1, b)/(\gamma(a_1, b) + \gamma(a_2, b))$ suit une loi bêta de paramètres a_1 et a_2.

2. La variable aléatoire $\beta(1, 1)$ est uniformément répartie sur l'intervalle [0, 1] (cf. n° 6.1.7).

3. La fonction de répartition du carré de la variable aléatoire de Student $t^2(m)$ (cf. n° 6.2.2) est liée à la fonction de répartition de la variable aléatoire β par la relation

$$F_{t^2(m)}(x) = F_{\beta\left(\frac{1}{2}, \frac{m}{2}\right)}\left(\frac{x^2}{m+x^2}\right).$$

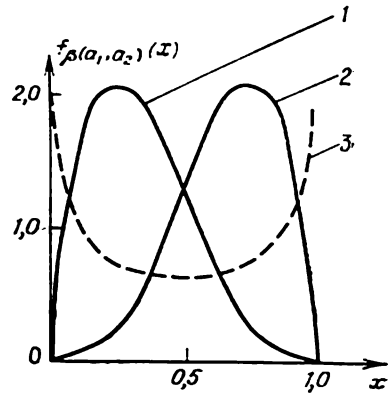


Fig. 6.5. Représentations graphiques des densités de la loi bêta pour différentes valeurs des paramètres a_1 et a_2: *1* — $a_1 = 2$, $a_2 = 4$; *2* — $a_1 = 4$, $a_2 = 2$; *3* — $a_1 = 1/2$, $a_2 = 1/2$

4. La fonction de répartition d'une variable aléatoire $F(m_1, m_2)$ (cf. n° 6.2.3) est liée à celle de la variable aléatoire β par la relation

$$F_{F(m_1, m_2)}(x) = F_{\beta\left(\frac{m_1}{2}, \frac{m_2}{2}\right)}\left(\frac{m_1 x}{m_2 + m_1 x}\right).$$

5. Entre la fonction de répartition de la variable aléatoire β et celles des lois binomiales (cf. n° 6.1.1), on a les relations suivantes:

$$F_{\beta(m, n-m+1)}(x) = \sum_{k=m}^{n} C_n^k x^k (1-x)^{n-k};$$

$$F_{\beta(n, m)}(1-x) = \sum_{k=n}^{\infty} C_{m+k-1}^{m-1} x^m (1-x)^k.$$

6. Une analyse directe de la densité (6.29) met en évidence la symétrie des densités $f_{\beta(a_1, a_2)}(x)$ et $f_{\beta(a_2, a_1)}(x)$ par rapport à la droite $x = 0{,}5$ (cf. fig. 6.5), ce qui en termes de fonctions de répartition s'écrit

$$F_{\beta(a_1, a_2)}(x) = 1 - F_{\beta(a_2, a_1)}(1-x)$$

(c'est pourquoi, notamment, on se limite généralement au cas $0 < a_1 \leqslant a_2$ lorsqu'on compose les tables de la loi bêta).

Les principales caractéristiques numériques de la variable aléatoire $\beta(a_1, a_2)$ sont :

la *moyenne* : $\mathbf{E}\beta(a_1, a_2) = \dfrac{a_1}{a_1+a_2}$;

le *mode* : $x_{\text{mod}} = \dfrac{a_1-1}{a_1+a_2-2}$ (pour $a_1 > 1$ et $a_2 > 1$) ;

la *variance* : $\mathbf{V}\beta(a_1, a_2) = \dfrac{a_1 a_2}{(a_1+a_2)^2(a_1+a_2+1)}$;

le *coefficient de dissymétrie* : $\beta_1 = \dfrac{2(a_2-a_1)\sqrt{a_1+a_2+1}}{(a_1+a_2+2)\sqrt{a_1 a_2}}$.

6.3. Technique de simulation statistique d'observations suivant une loi donnée

6.3.1. Génération de nombres aléatoires uniformément répartis sur l'intervalle [0, 1]. Il existe deux méthodes de génération de nombres aléatoires par un ordinateur : une méthode « physique » qui consiste à connecter l'ordinateur à un générateur de nombres aléatoires (par exemple, un compteur déterminant le nombre de particules α émises par une source radioactive durant un intervalle de temps donné) et une méthode mathématique qui consiste à utiliser les commandes standards de l'ordinateur pour générer une *suite régulière* de nombres qui semblent *aléatoires* pour un observateur et qui vérifient toutes les inégalités que sont censés vérifier les nombres aléatoires. De tels nombres sont dits *pseudo-aléatoires*. La méthode mathématique fait actuellement recette pour plusieurs raisons. Tout d'abord en simulation statistique il est important de pouvoir reproduire une suite de nombres aléatoires pour voir, par exemple, comment une autre méthode de traitement statistique opérera avec les mêmes données. D'autre part, il est difficile de garantir un fonctionnement normal des générateurs physiques. Et, enfin, il existe actuellement des générateurs mathématiques qui sont simples et sûrs à la fois.

Pour générer des nombres pseudo-aléatoires $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ on se sert le plus souvent de la *méthode des résidus* (générateur multiplicatif) :

u_0 est un entier $\leqslant 2^m$,

$$u_n \equiv u_{n-1} \cdot M \pmod{2^m}, \qquad \xi_n = u_n \cdot 2^{-m}, \tag{6.30}$$

où m et M sont des entiers constants spécialement choisis. La fixation du nombre initial u_0 définit une seule suite de nombres ξ_i. Comme le nombre des u_n est inférieur à celui des résidus mod 2^m, la suite ξ_i

admet une période $L \leqslant 2^m$. Cette période dépend en générale de u_0. La formule (6.30) peut être mise sous la forme

$$\xi_0 = u_0 2^{-m}, \ \xi_n = \{M\xi_{n-1}\}, \tag{6.31}$$

où $\{V\}$ représente la partie fractionnaire du nombre V.

La formule (6.31) permet de se représenter plus nettement le caractère de la dépendance entre les ξ_i. Développons à cet effet le nombre ξ_0 sous forme d'une fraction infinie suivant les puissances de M^{-1}, c'est-à-dire que $\xi_0 = a_1M^{-1} + a_2M^{-2} + a_3M^{-3} + \ldots$ ou encore $\xi_0 = 0, a_1a_2a_3 \ldots$, où chaque a_i peut prendre les valeurs de 0 à $M - 1$. De la formule (6.31), il s'ensuit alors que $\xi_n = 0, a_na_{n+1} \ldots$, c'est-à-dire que pour obtenir ξ_n il suffit de déplacer dans ξ_0 la virgule de n rangs à droite et d'éliminer la partie entière égale à $a_1a_2 \ldots a_{n-1}$.

Les valeurs suivantes de m et M ont fourni de bons résultats pratiques :

$$m = 2^{36}, \quad M = 5^{15} \quad [51];$$
$$m = 2^{40}, \quad M = 5^{17} \quad [51];$$
$$m = 2^{35}, \quad M = 5^{51} \quad [79];$$
$$m = 2^{35}, \quad M = 2718281821 \quad [79].$$

La qualité d'une suite de nombres pseudo-aléatoires se mesure à l'uniformité avec laquelle les vecteurs de dimension k pour $k = 1, 2, \ldots, K$, où K est de l'ordre de plusieurs dizaines,

$$\eta_1^T = (\xi_1, \ldots, \xi_k); \quad \eta_2^T = (\xi_{k+1}, \ldots, \xi_{2k}); \quad \eta_n^T = (\xi_{k(n-1)+1}, \ldots, \xi_{nk})$$

recouvrent le cube unité de dimension k.

Dans [51] on peut trouver les résultats de la vérification d'un générateur pour l'ordinateur BESM-6. Les problèmes mathématiques relevant de la construction des générateurs de nombres pseudo-aléatoires sont traités le plus systématiquement et le plus exhaustivement dans le tome 2 de la monographie de Knuth [79]. On y décrit aussi une méthode montrant comment utiliser deux générateurs « indépendants » de périodes L_1 et L_2 pour obtenir un générateur d'une période de l'ordre de $L_1 \cdot L_2$. Cette méthode peut être utile pour générer des nombres pseudo-aléatoires sur un ordinateur à mots courts.

Pour décrire les méthodes de génération des nombres aléatoires, on se servira dans la suite du langage Fortran et l'on admettra que dans le programme principal est décrit un générateur de nombres pseudo-aléatoires qu'on peut appeler à l'aide de l'opérateur CALL RAND (R), où R est le nom du nombre pseudo-aléatoire. Le sous-programme RAND (R) calcule le nombre pseudo-aléatoire ξ_{n+1} à l'aide du nombre ξ_n contenu dans la cellule R et le loge dans R.

6.3.2. Simulation des variables aléatoires discrètes. Méthode standard. La méthode générale de simulation d'une variable aléatoire discrète X prenant des valeurs x_k avec les probabilités p_k, $k = 1, \ldots, m$; $\sum_k p_k = 1$, est basée sur la formule évidente suivante:

$$P\{X = x_l\} = P\left\{\sum_{k=0}^{l-1} p_k < \xi \leqslant \sum_{k=0}^{l} p_k\right\} = p_l,$$

où, pour la commodité de l'écriture, on a posé $p_0 = 0$, et ξ est une variable aléatoire uniformément répartie sur l'intervalle [0, 1].

Supposons que les valeurs x_k et p_k sont enregistrées respectivement dans les cellules A_k et P_k, $k = 1, \ldots, K$. Le programme (en Fortran) de calcul de la variable aléatoire X s'écrit dans ce cas:

```
SUBROUTINE DISCR (K, A, P, X)
  1. DIMENSION A (K), P (K)
  2. CALL RAND (R)
  3. DO 8 I = 1, K
  4. R = R — P (I)
  5. IF (R) 6, 6, 8
  6. X = X (I)
  7. GO TO 9
  8. CONTINUE
  9. RETURN
 10. END
```

Si les probabilités p_k sont liées par une relation récurrentielle simple $p_{k+1} = p_k r(k)$, on peut ne pas introduire le bloc P à l'avance, mais calculer les valeurs p_k dans le programme. Par exemple:

a. Pour la loi binomiale de paramètres p, n

$$p_k = C_n^k p^k (1-p)^{n-k} \text{ et } r(k) = \frac{p_{k+1}}{p_k} = \frac{(n-k)}{k+1} \cdot \frac{p}{1-p}; \quad k = 0, 1, \ldots, n-1.$$

b. Pour la loi de Poisson de paramètre λ

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda} \text{ et } r(k) = \lambda/(k+1); \quad k = 0, 1, \ldots, \ldots$$

a. *Loi binomiale.* Pour calculer les valeurs d'une variable aléatoire binomiale X de paramètres p et n, on peut se servir aussi de la simulation statistique, plus exactement, procéder à n réalisations indépendantes d'une variable aléatoire uniformément répartie ξ_i, $i = 1, \ldots, n$, et poser X égal au nombre de cas où $\xi_i < p$. Le

programme correspondant (en Fortran) s'écrit :

```
SUBROUTINE BINOM (N, P, X)
  1. X = 0.0
  2. DO 6 I = 1, N
  3. CALL RAND (R)
  4. IF (R — P) 5, 5, 6
  5. X = X + 1.0
  6. CONTINUE
  7. RETURN
  8. END
```

b. *Loi de Poisson.* Les valeurs empiriques de la variable aléatoire poissonnienne X de paramètre λ sont générées aussi par une méthode de simulation statistique directe basée sur le fait que X peut être définie comme

$$X = \min \{n : \prod_{i=1}^{n} \xi_i < \exp\{-\lambda\}\}.$$

Le programme correspondant s'écrit :

```
SUBROUTINE POISSN (P, X)
  1. X = 0.0
  2. B = EXP {—P}
  3. TR = 1.0
  4. CALL RAND (R)
  5. TR = TR*R
  6. IF (TR — B) 9, 7, 7
  7. X = X + 1.0
  8. GO TO 4
  9. RETURN
 10. END
```

EXP (X) appelle la procédure standard de calcul de exp $\{x\}$. Le paramètre formel p remplace λ.

6.3.3. Simulation de lois continues. Commençons par une *méthode standard.* Supposons qu'une variable aléatoire η admet une fonction de répartition $F(x)$. La méthode standard de simulation de η repose sur le fait que la variable aléatoire $F(\eta)$ est uniformément répartie, donc, la variable aléatoire $F^{-1}(\xi)$, où $F^{-1}(\cdot)$ est la fonction inverse de $F(x)$, est répartie comme η.

A titre d'exemple traitons le cas où η suit une *loi exponentielle*, c'est-à-dire que $P\{\eta < x\} = 1 - \exp\{-lx\}$. Alors $F^{-1}(u) =$

$= -\ln(1-u)/l$ et η peut être simulé comme

$$\eta = F^{-1}(\xi) = -\ln(1-\xi)/l$$

ou comme $-\ln \xi/l$, puisque ξ et $1-\xi$ sont parentes.

Le programme correspondant est

```
SUBROUTINE EXP (L, X)
  1. CALL RAND (R)
  2. X = — LN (R)/L
  3. RETURN
  4. END
```

Simulation d'une variable aléatoire normale à une dimension. D'après le théorème limite central (cf. n° 7.3.1), la variable aléatoire

$$\eta(n) = \sqrt{12/n} \sum_{i=1}^{n} (\xi_i - 0{,}5) \tag{6.32}$$

suit approximativement la loi normale réduite.

Le programme correspondant est le suivant:

```
SUBROUTINE NORMAL (EX, STDX, X)
  1. SUM = 0.0
  2. DO 4 I = 1, 12
  3. CALL RAND (R)
  4. SUM = SUM + R
  5. X = STDX*(SUM — 6) + EX
  6. RETURN
  7. END
```

Lorsque $n = 12$, la formule (6.32) est souvent utilisée pour simuler la loi normale dans le cas où les valeurs élevées (>3) de $|\eta|$ ne jouent pas un rôle essentiel. Pour améliorer l'approximation, L. Bolchev [26] a proposé une correction non linéaire pour $\eta(n)$:

$$\zeta(n) = \eta(n) + (\eta^3(n) - 3\eta(n))/(20n).$$

Mais si l'analyste s'intéresse aux grandes valeurs de $|\eta|$ ou s'il a besoin de réaliser plusieurs fois la loi normale, il peut se servir des formules exactes qui mettent en jeu un nombre moindre de nombres pseudo-aléatoires. Dans ce cas, les variables aléatoires normales réduites sont générées par couples:

$$\eta_1 = \sqrt{-2\ln \xi_1}\sin(2\pi\xi_2) \quad \text{et} \quad \eta_2 = \sqrt{-2\ln \xi_1}\cos(2\pi\xi_2). \tag{6.33}$$

Les formules (6.33) sont basées sur une propriété classique de la loi normale, savoir que si η_1 et η_2 sont des variables aléatoires réduites

indépendantes, l'angle formé par l'axe des abscisses et le vecteur de sommet (η_1, η_2) suit une loi uniforme et ne dépend pas de $\sqrt{\eta_1^2 + \eta_2^2}$. Le carré de la longueur du vecteur (η_1, η_2) est réparti dans ce cas suivant une loi de χ^2 à deux degrés de liberté qui est simulée comme un cas particulier de la loi exponentielle avec $l = 0{,}5$.

Simulation d'un vecteur normal (M, Σ) *non dégénéré multidimensionnel.* Tout d'abord, on simule le vecteur $\eta' = (\eta_1, \ldots, \eta_p)$, où η_i $(i = 1, \ldots, p)$ sont des variables aléatoires normales réduites indépendantes, à l'aide d'une des méthodes décrites plus haut, puis on calcule le vecteur ζ en se servant de la transformation $\zeta = A\eta + M$, où A est une matrice triangulaire telle que $AA' = \Sigma$.

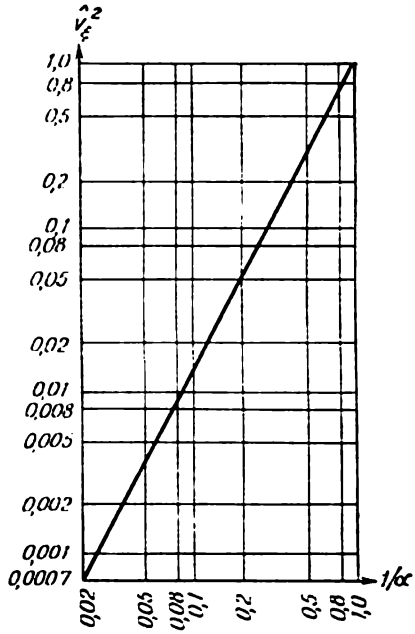


Fig. 6.6. Abaque pour la détermination d'une estimation du paramètre de forme α dans la loi de Weibull sur le vu du carré du coefficient empirique de variation $\hat{V}\xi^2$

Conclusions

Dans le tableau 6.3 sont reproduites les propriétés des lois les plus couramment utilisées dans les recherches statistiques. Ce tableau résume le présent chapitre. Pour s'en servir, il faut savoir que

la colonne 3 qui est destinée à recevoir la procédure générale (ou le mécanisme) de formation des valeurs de la variable aléatoire ne décrit qu'*une des éventuelles variantes* de ce mécanisme;

la colonne 6 (consacrée aux estimations statistiques des paramètres de la loi de probabilité) ne doit être consultée qu'après lecture du chapitre 8 qui permet de se faire une idée sur la méthode d'établissement des estimations citées;

dans ce tableau on n'indique une méthode d'acquisition des estimations que dans le cas où la méthode du maximum de vraisemblance n'a pas été utilisée;

dans la colonne 4, la liste des exemples de situations réelles justiciables de la loi de probabilité indiquée est loin d'être complète et sa description n'a pas l'ambition d'être circonstanciée et rigoureuse.

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
1	Binomiale $\nu_p(m)$	Nombre $\nu_p(m)$ d'apparitions d'un événement dans une suite de m expériences indépendantes, la probabilité d'apparition de cet événement dans une épreuve étant constante et égale à p	1. Nombre de pièces défectueuses dans un lot de taille m extrait d'une production stationnaire en série 2. Nombre d'individus (d'objets) possédant certains caractères parmi m prélevés au hasard dans une population générale
2	Binomiale négative $\nu_p^-(k)$	Nombre $\nu_p^-(k)$ d'expériences indépendantes nécessaires pour réaliser k fois l'événement étudié, la probabilité d'apparition de cet événement étant constante et égale à p	1. Durée de vie d'un système (mesurée en cycles de fonctionnement) possédant $k-1$ éléments en réserve (se branchant automatiquement) 2. Taille de l'échantillon nécessaire pour prélever au hasard k objets jouissant des propriétés données
3	Hypergéométrique $\nu_{M,N}(m)$	Nombre $\nu_{M,N}(m)$ d'objets possédant une propriété donnée parmi m objets tirés au hasard (sans remise) dans une population générale de N objets dont M jouissent de la dite propriété	Nombre de pièces défectueuses dans un échantillon de taille m prélevé au hasard dans un lot de N pièces dont M sont défectueuses

Tableau 6.3

Définition analytique de la loi : polygone de fréquence $P\{\xi = x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$P\{\nu_p(m) = x\} = C_m^x p^x (1-p)^{m-x}$, $x = 0, 1, 2, \ldots, m$	$\hat{p} = \dfrac{\sum_{i=1}^{n} x_i}{mn}$, où x_i est le nombre d'apparitions de l'événement étudié dans la i-ième observation, c'est-à-dire dans la i-ième série de m épreuves indépendantes	$E\nu_p(m) = mp$ $V\nu_p(m) = mp(1-p)$ $\beta_1 = \dfrac{1-2p}{\sqrt{mp(1-p)}}$ $\beta_2 = \dfrac{1-6p(1-p)}{mp(1-p)}$
$P\{\nu_p^-(k) = x\} = C_{x-1}^{k-1} p^k (1-p)^{x-k}$, $x = k, k+1, k+2, \ldots$	$\hat{p} = \dfrac{k}{n} \sum_{i=1}^{n} \dfrac{1}{x_i}$, où x_i est le nombre d'expériences réalisées au cours de la i-ième observation jusqu'à la k-ième apparition de l'événement étudié	$E\nu_p^-(k) = \dfrac{k}{p}$ $V\nu_p^-(k) = \dfrac{k(1-p)}{p^2}$ $\beta_1 = \dfrac{2-p}{\sqrt{k(1-p)}}$ $\beta_2 = \dfrac{1+4(1-p)}{k(1-p)} + \dfrac{(1-p)^2}{k(1-p)}$
$P\{\nu_{M,N}(m) = x\} = \dfrac{C_M^x C_{N-M}^{m-x}}{C_N^m}$, $x = \max\{0, m-(N-M)\}, \max\{0, m-(N-M)\}+1, \ldots, \min\{m, M\}$	Pour une valeur connue du paramètre N : $\hat{M} = \dfrac{N}{m} \cdot \dfrac{\sum_{i=1}^{n} x_i}{n}$, où x_i est le nombre d'objets dénombrés au cours de la i-ième observation, possédant la propriété indiquée parmi m objets prélevés au hasard dans une population de taille N	$E\nu_{M,N}(m) = m\dfrac{M}{N}$ $V\nu_{M,N}(m) = m\dfrac{M}{N-1}\left(1-\dfrac{M}{N}\right) \times \left(1-\dfrac{m}{N}\right)$ $\beta_1 = \dfrac{\left(1-2\dfrac{M}{N}\right)}{\sqrt{m\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)}} \times \dfrac{(N-2m)\sqrt{n-1}}{(N-2)\sqrt{N-m}}$ $\beta_2 =$ (cf. n° 6.1.2)

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
4	Poisson (des événements rares) ν_0	Nombre d'apparitions pendant une unité de temps d'un événement dont la réalisation ne dédend ni du nombre ni de la date de ses réalisations passées et n'influe pas sur ses réalisations futures, les expériences ayant lieu en régime stationnaire	1. Nombre de pannes d'un processus industriel (c'est-à-dire fonctionnant en régime d'exploitation normale) en une unité de temps 2. Nombre d'unités arrivant pendant une unité de temps dans un système de files d'attente 3. Nombre de sinistres (ou de décès des suites de maladies rares) pendant une unité de temps dans une population donnée
5	Polynomiale (multinomiale) $\nu_p(m) = (\nu_p^{(1)}(m), \ldots, \nu_p^{(l)}(m))$	Nombres $\nu_p^{(1)}(m)$, $\nu_p^{(2)}(m), \ldots, \nu_p^{(l)}(m)$ d'apparitions d'événements $A_1, A_2, \ldots, A_l$ (formant un système complet, c'est-à-dire deux à deux incompatibles et dont la somme est un événement sûr) dans une série de m expériences indépendantes, la probabilité de A_j étant constante et égale à p_j au cours d'une épreuve	Collection de nombres définissant la distribution de m objets en fonction de l caractères, ces objets étant extraits d'une population générale dont chaque objet possède obligatoirement l'un de ces caractères
6	Normale (de Gauss) $\xi(a, \sigma^2)$	Les valeurs de la variable aléatoire se forment sous l'influence d'un grand nombre de facteurs aléatoires mutuellement indépendants, l'action de chacun d'eux étant petite et ne prédominant pas celle des autres, et le caractère de cette action	1. Ecarts par rapport aux valeurs nominales des paramètres d'articles produits en série dans des conditions stationnaires 2. Erreur de calcul 3. Erreur de tir

Suite

Définition analytique de la loi : polygone de fréquence $P\{\xi=x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$P\{\nu_0=x\}=\frac{\lambda^x}{x!}e^{-\lambda}$, $x=0, 1, 2, \ldots$	$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n} x_i$, où x_i est le nombre d'apparitions de l'événement étudié pendant la i-ième unité de temps	$E\nu_0=\lambda$ $V\nu_0=\lambda$ $\beta_1=\frac{1}{\sqrt{\lambda}}$ $\beta_2=\frac{1}{\lambda}$
$P\{\nu_p^{(1)}(m)=x^{(1)}, \ldots, \nu_p^{(l)}(m)=x^{(l)}\}=\frac{m!}{x^{(1)}!\ldots x^{(l)}!}\times p_1^{x^{(1)}}\ldots p_l^{x^{(l)}}$, $x^{(j)}=0, 1, 2, \ldots, m$ $(j=1, 2, \ldots, l)$	$\hat{p}^{(j)}=\frac{\sum_{i=1}^{n} x_i^{(j)}}{mn}$, où $x_i^{(j)}$ est le nombre d'apparitions de l'événement A_j au cours de la i-ième observation, c'est-à-dire dans la i-ième série de m épreuves indépendantes	$E\nu_p^{(j)}(m)=mp_j$, $j=1, 2, \ldots, l$ $V\nu_p^{(j)}(m)=mp_j\times(1-p_j)$, $j=1, \ldots, l$ $\sigma_{jk}=-mp_jp_k$, $j, k=1, 2, \ldots, l$ $j\neq k$
Dimension un : $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-a)^2}$, $-\infty<x<\infty$ Dimension n :	$\hat{a}=\frac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{a})^2$ $\hat{M}_1=\frac{1}{n}\sum_{i=1}^{n} X_i$	$E\xi(a, \sigma^2)=a$ $V\xi(a, \sigma^2)=\sigma^2$ $\beta_1=0$ $\beta_2=0$ $E\{(\xi^{(j)}-m_1^{(j)})\times(\xi^{(k)}-m_1^{(k)})\}=\sigma_{jk}$

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
		étant additif, c'est-à-dire que l'action d'un facteur aléatoire sur une quantité a nous donne une quantité $a+\Delta$, où Δ est une contribution aléatoire relativement petite et de signe équiprobable	
7	Lognormale η	Les valeurs de la variable aléatoire se forment sous l'action d'un grand nombre de facteurs aléatoires mutuellement indépendants, l'action de chacun d'eux étant uniformément insignifiante et le caractère de cette action, multiplicatif : l'« accroissement » aléatoire dû à l'action de chaque facteur est proportionnel à la valeur de la variable aléatoire au moment de cette action, autrement dit, l'action d'un facteur sur une quantité a nous donne une quantité $a+\Delta a$, où Δ est un petit coefficient aléatoire de signe équiprobable	1. Le salaire d'un employé prélevé dans une population générale 2. Le revenu moyen par membre d'une famille prélevée dans une population générale de familles 3. La durée de vie d'une pièce exploitée en régime d'usure et de vieillissement 4. Dimensions et volume de particules désagrégées

Suite

Définition analytique de la loi : polygone de fréquence $P\{\xi = x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \lvert\Sigma\rvert^{\frac{1}{2}}} \times$ $\times e^{-\frac{1}{2}(X-M_1)'\Sigma^{-1}(X-M_1)}$, $-\infty < x^{(i)} < \infty$, $i = 1, 2, \ldots, p$, où $X = (x^{(1)}, \ldots, x^{(p)})'$ est le vecteur colonne des valeurs données (courantes) des composantes de la variable aléatoire multidimensionnelle	$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{M}_1) \times$ $\times (X_i - \hat{M}_1)'$, où $\hat{M}_1 = (m_1^{(1)}, \ldots$ $\ldots, m_1^{(p)})'$ et $\hat{\Sigma} = (\hat{\sigma}_{ik})_j$, $k = 1, \ldots \overline{p}$	$(j, k = 1, 2, \ldots, p)$ $m_1^{(l)} = E\xi^{(l)}$ $(l = 1, 2, \ldots p)$
$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} \times$ $\times e^{-\frac{(\ln x - \ln a)^2}{2\sigma^2}}$, $0 < x < \infty$	$\hat{a} = \hat{x}_{\text{méd}}(n)$ (cf. n°s 5.6.2 et 5.6.4); $\hat{\sigma}^2 = \frac{1}{n-1} \times$ $\times \sum_{i=1}^{n} (\ln x_i - \ln \hat{a})^2$	$E\eta = ae^{\frac{1}{2}\sigma^2}$ $V\eta = a^2 e^{\sigma^2}(e^{\sigma^2} - 1)$ $\beta_1 = (e^{\sigma^2} - 1)^{\frac{1}{2}} \times$ $\times (e^{\sigma^2} + 2)$ $\beta_2 = (e^{\sigma^2} - 1) \times$ $\times (e^{3\sigma^2} + 3e^{2\sigma^2} +$ $+ 6e^{\sigma^2} + 6)$

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
8	Uniforme (rectangulaire) ξ	Le mécanisme de formation des valeurs de la variable aléatoire est tel que la probabilité d'obtenir une observation dans n'importe quel voisinage de l'intervalle $[a, b]$ de ses valeurs possibles dépend uniquement de la largeur de ce voisinage et non pas de sa position dans $[a, b]$	1. Erreur d'arrondi dans des calculs numériques où le nombre de décimales est fixé 2. Durée d'attente d'un service dans un système dont les guichets (stations) fonctionnent périodiquement et dont les unités (clients) arrivent uniformément pendant cet intervalle 3. Approximation zéro dans la description de la répartition *a priori* des paramètres analysés dans l'approche bayésienne (cf. n° 8.6.6)
9	Exponentielle ξ	Intervalle de temps aléatoire séparant deux événements poissonniens (cf. n° 4 de ce tableau)	1. Durée de service dans des systèmes de files d'attente 2. Durée de vie d'une pièce fonctionnant en régime normal d'exploitation (cf. n° 6.1.8) 3. Durée d'une opération technologique 4. Intervalle de temps séparant deux pannes consécutives de machines fonctionnant en régime stationnaire

Suite

Définition analytique de la loi : polygone de fréquence $P(\xi = x)$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a, b]; \\ 0, x \notin [a, b], \end{cases}$ $x \in [a, b]$	$\hat{a} = x_{\min}(n) - \dfrac{x_{\max}(n) - x_{\min}(n)}{n-1}$ $\hat{b} = x_{\max}(n) + \dfrac{x_{\max}(n) - x_{\min}(n)}{n-1}$ où $x_{\min}(n) = \min\{x_1, x_2, \ldots, x_n\}$; $x_{\max}(n) = \max\{x_1, x_2, \ldots, x_n\}$	$E\xi = \dfrac{a+b}{2}$ $V\xi = \dfrac{(b-a)^2}{12}$ $\beta_1 = 0$ $\beta_2 = -1{,}2$
$f(x) = \begin{cases} \lambda e^{-\lambda(x-\theta)} \\ \text{pour } x \geqslant \theta; \\ 0 \text{ pour } x < \theta, \end{cases}$ $x \in [\theta, \infty[$	$\hat{\theta} = x_{\min}(n) - \dfrac{\bar{x} - x_{\min}(n)}{n}$; $\lambda = \dfrac{1}{\bar{x} - \hat{\theta}}$, où $x_{\min}(n) = \min\{x_1, x_2, \ldots, x_n\}$; $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$	$E\xi = \dfrac{1}{\lambda} + \theta$ $V\xi = \dfrac{1}{\lambda^2}$ $\beta_1 = 2$ $\beta_2 = 6$

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
10	Weibull (troisième type de lois limites des valeurs empiriques extrêmes) ξ	*Première variante :* les observations suivant la loi de Weibull sont générées par des essais « à la durée de vie », la fonction d'intensité des défaillances « le quotient de mortalité ») $\lambda(t)$ étant une fonction puissance, c'est-à-dire que $\lambda(t) = ct^{\alpha-1}$ $(\alpha > 0)$ *Deuxième variante* : la loi de Weibull décrit la répartition des plus petites valeurs dans des échantillons de grande taille extraits de populations générales, l'intervalle des valeurs possibles de ξ étant borné à gauche	1. Durée de vie d'un système ou d'une pièce fonctionnant en régime de rodage $(0 < \alpha < 1)$, d'exploitation normale $(\alpha = 1)$ ou d'usure et de vieillissement $(\alpha > 1)$ 2. Nombre de cycles (durée) avant la destruction d'un prototype dans des essais à la fatigue 3. Minimum (sur plusieurs années) d'un produit consommé en une année (par exemple, consommation annuelle d'eau sur plusieurs années dans l'analyse des sécheresses)

Suite

Définition analytique de la loi : polygone de fréquence $P\{\xi = x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \begin{cases} \lambda_0 \alpha x^{\alpha-1} \times \\ \times e^{-\lambda_0 x^{\alpha-1}} \\ \text{pour } x \geqslant 0; \\ 0 \text{ pour } x < 0, \end{cases}$ $x \in [0, \infty[$	L'estimation $\hat{\alpha}$ du paramètre α est définie tout d'abord comme la racine de l'équation $1+\hat{v}_\xi^2 = \dfrac{\Gamma\left(1+\frac{2}{\hat{\alpha}}\right)}{\Gamma^2\left(1+\frac{1}{\hat{\alpha}}\right)}$, où $\hat{v}_\xi^2$ est le coefficient de variation empirique de ξ (cf. n° 5.6.3, ainsi que la figure 6.6), puis est calculée l'estimation $\hat{\lambda} = \left(\dfrac{\Gamma\left(1+\frac{1}{\hat{\alpha}}\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i}\right)^{\hat{\alpha}}$ (méthode des moments)	$E\xi = \lambda_0^{-\frac{1}{\alpha}} \Gamma\left(1+\frac{1}{\alpha}\right)$ $V\xi = \lambda_0^{-\frac{2}{\alpha}} \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$ $\beta_1 = \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^{-3} \times \left[\Gamma\left(1+\frac{3}{\alpha}\right) - 3 \times \Gamma\left(1+\frac{2}{\alpha}\right) \times \Gamma\left(1+\frac{1}{\alpha}\right) + 2 \times \Gamma^3\left(1+\frac{1}{\alpha}\right)\right]$ $\beta_2 = \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^{-4} \times \left[\Gamma\left(1+\frac{4}{\alpha}\right) - 4 \times \Gamma\left(1+\frac{3}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right) + 6\Gamma\left(1+\frac{2}{\alpha}\right) \times \Gamma^2\left(1+\frac{1}{\alpha}\right) - 3\Gamma^4\left(1+\frac{1}{\alpha}\right)\right] - 3$

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
11	Laplace (exponentielle bilatérale) ξ		Répartition de la composante aléatoire résiduelle ε dans certains modèles de régression de la forme $Y=f(X;\ \theta)+\varepsilon$
12	Pareto ξ	Caractère de la variation aléatoire de la variable aléatoire étudiée dans des populations générales d'où ont été extraits tous les éléments dont les valeurs du caractère étudié sont inférieures à un niveau c_0	Répartition du revenu moyen par personne dans une famille extraite d'une population composée uniquement de familles dont le revenu moyen par membre est supérieur à un niveau donné c_0

Suite

Définition analytique de la loi : polygone de fréquence $P\{\xi = x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \frac{\lambda}{2} e^{-\lambda \lvert x \rvert}$, $-\infty < x < \infty$	$\hat{\lambda} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \lvert x_i \rvert}$	$E\xi = 0$ $V\xi = \frac{2}{\lambda^2}$ $\beta_1 = 0$ $\beta_2 = 3$
$f(x) = \begin{cases} \frac{\alpha}{c_0} \left(\frac{c_0}{x}\right)^{\alpha+1} \text{ pour } x \geqslant c_0; \\ 0 \text{ pour } x < c_0, \end{cases}$ $c_0 \leqslant x < \infty$	1. Méthode des moments pour $\alpha > 2$ (cf. n° 8.6.2) : $\begin{cases} \hat{\alpha} = 1 + \sqrt{1 + \hat{v}_\xi^{-2}}; \\ \hat{c} = \left(1 - \frac{1}{\hat{\alpha}}\right) \times \\ \times \frac{1}{n} \sum_{i=1}^{n} x_i \end{cases}$ où $\hat{v}_\xi$ est le coefficient de variation empirique 2. Réalisation approchée de la méthode du maximum de vraisemblance (cf. n° 8.6.1) : $\hat{c}_0 = x_{\min}(n) - \frac{1}{n}$; $\hat{\alpha} = \left[\frac{1}{n} \times \right.$ $\left. \times \sum_{i=1}^{n} \ln\left(\frac{x_i}{\hat{c}_0}\right)\right]^{-1}$, où $x_{\min}(n) =$ $= \min\{x_1, \ldots, x_n\}$	$E\xi = \frac{\alpha}{\alpha - 1} c_0$ (existe pour $\alpha > 1$) $V\xi = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} c_0^2$ (existe pour $\alpha > 2$) $E\xi^k = \frac{\alpha}{\alpha - k} c_0^k$, $k = 3, 4, \ldots$ (existe pour $\alpha > k$)

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
13	Cauchy ξ		Le rapport de deux variables aléatoires normales indépendantes de moyenne $a=0$ et de variance $\sigma^2=1$ suit une loi de Cauchy de paramètres $a=0$ et $c=1$
14	χ^2 à m degrés de liberté $\chi^2(m)$	Somme des carrés de m variables aléatoires normales réduites indépendantes $\xi_1(0,1), \xi_2(0,1), \ldots, \xi_m(0,1)$: $\chi^2(m)=\sum_{i=1}^{m}\xi_i^2(0,1)$	1. Variance empirique normalisée construite sur un échantillon d'une population générale normale (cf. (6.19)) 2. Mesure de l'écart entre les lois théorique et empirique (cf. (6.20))
15	Student à m degrés de liberté $t(m)$	$t(m)=\dfrac{\xi_0(0,\sigma^2)}{\left(\frac{1}{m}\sum_{i=1}^{m}\xi_i^2(0,\sigma^2)\right)^{\frac{1}{2}}}$, où $\xi_0(0,\sigma^2), \ldots, \xi_m(0,\sigma^2)$ sont des variables aléa-	1. Mesure normalisée de l'écart de deux moyennes empiriques calculées sur deux échantillons indépendants d'une population générale normale (cf. (6.24)) 2. Ecart normalisé de la moyenne empirique (construite sur un échantillon d'une population nor-

Suite

Définition analytique de la loi : polygone de fréquence $P(\xi = x)$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \frac{1}{\pi} \frac{c}{c^2 + (x-a)^2}$, $-\infty < x < \infty$	Méthode de substitution, dans les relations entre quantiles et paramètres de la loi, des quantiles empiriques : $\hat{a} = x_{\text{méd}}(n)$; $\hat{c} = x_{([0,75n+1])} - x_{\text{méd}}(n)$, où $[A]$ est la partie entière de A ; $x_{\text{méd}}(n)$ est la médiane empirique, $x_{([0,75n+1])}$ le quantile empirique d'ordre 0,75 (c'est-à-dire le $[0,75n+1]$-ième variationnelle construite sur l'échantillon étudié (cf. n°s 5.6.4 et 5.6.5))	
$f(x) = \frac{1}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} \times x^{\frac{m}{2}-1} e^{-\frac{x}{2}}$, $0 \leqslant x < \infty$	$\hat{m} = \left[\frac{1}{n} \sum_{i=1}^{n} x_i\right]$, où $[A]$ est la partie entière de A (méthode des moments)	$E\chi^2(m) = m$ $V\chi^2(m) = 2m$ $\beta_1 = \frac{2^{\frac{3}{2}}}{\sqrt{m}}$ $\beta_2 = \frac{12}{m}$
$f(x) = \frac{1}{\sqrt{\pi m}} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \times \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}$, $-\infty < x < \infty$	$\hat{m} = \left[\frac{2s^2}{s^2 - 1}\right]$, où $[A]$ est la partie entière de A et $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, la variance empirique (méthode des moments)	$Et(m) = 0$ $Vt(m) = \frac{m}{m-2}$ (existe pour $m > 2$) $\beta_1 = 0$ $\beta_2 = \frac{6}{m-4}$ (existe pour $m > 4$)

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
		toires normales indépendantes de moyenne 0 et de variance σ^2 (σ^2 est un nombre strictement positif quelconque)	male) par rapport à la valeur théorique respective (cf. (6.23))
16	Fisher (loi F à (m_1, m_2) degrés de liberté) $F(m_1, m_2)$	$F(m_1, m_2) = \dfrac{\dfrac{1}{m_1}\sum_{i=1}^{m_1} \xi_i^2(0, \sigma^2)}{\dfrac{1}{m_2}\sum_{i=m_1+1}^{m_1+m_2} \xi_i^2(0, \sigma^2)}$ où $\xi_1(0, \sigma^2)$, $\xi_2(0, \sigma^2)$, ..., $\xi_{m_1+m_2}^2(0, \sigma^2)$ sont des variables aléatoires indépendantes normales de moyenne 0 et de variance σ^2 (σ^2 est un nombre quelconque > 0)	1. Rapport de deux variances empiriques construites sur deux échantillons indépendants extraits d'une population générale normale (cf. n° 6.2.3) 2. Analogue multidimensionnel de la statistique T^2 de Student qui décrit l'écart de deux moyennes vectorielles empiriques construites sur deux échantillons indépendants extraits d'une population générale normale multidimensionnelle (cf. (6.27))
17	Gamma $\gamma(a, b)$	—	—

Suite

Définition analytique de la loi : polygone de fréquence $P\{\xi = x\}$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f_F(m_1, m_2)(x) = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \times m_1^{\frac{m_1}{2}} m_2^{\frac{m_2}{2}} \times \frac{x^{\frac{m_1}{2}-1}}{(m_1x+m_2)^{\frac{m_1+m_2}{2}}}$, $0 \leqslant x < \infty$	Pour $m_2 > 4$: $\hat{m}_2 = \left[\frac{2\bar{x}}{\bar{x}-1}\right]$; $\hat{m}_1 = [2\hat{m}_2^2(\hat{m}_2 - 2)\{s^2(\hat{m}_2-2)^2(\hat{m}_2-4) - 2\hat{m}_2\}^{-1}]$, où $[A]$ est la partie entière de A et $\bar{x}$ et s^2 la moyenne et la variance empiriques (méthode des moments)	$EF(m_1, m_2) = \frac{m_2}{m_2-2}$ (existe pour $m_2 > 2$) $VF(m_1, m_2) = \frac{2m_2^2(m_1+m_2-2)}{m_1(m_2-2)^2(m_2-4)}$ (existe pour $m_2 > 4$) $\beta_1 = \frac{(2m_1+m_2-2)\times}{(m_2-6)\times} \frac{\times\sqrt{8(m_2-4)}}{\times\sqrt{m_1+m_2-2}}$ (existe pour $m_2 > 6$)
$f_\gamma(a, b)(x) = \begin{cases} \frac{b^a}{\Gamma(a)} \cdot x^{a-1}e^{-bx} \\ \text{pour } x \geqslant 0; \\ 0 \text{ pour } x < 0, \end{cases}$ $0 \leqslant x < \infty$	$\hat{a} = \left(\frac{\bar{x}}{s}\right)^2$, $\hat{b} = \frac{\bar{x}}{s^2}$, où $\bar{x}$ et s^2 sont la moyenne et la variance empiriques (méthode des moments)	$E\gamma(a, b) = \frac{a}{b}$ $V\gamma(a, b) = \frac{a}{b^2}$ $\beta_1 = \frac{2}{\sqrt{a}}$ $\beta_2 = \frac{6}{a}$

n°	Loi et variable aléatoire correspondante	Schéma général (mécanisme) de formation des variables aléatoires	Exemples de variables aléatoires suivant la loi considérée
18	Bêta $\beta(a_1, a_2)$	—	Vaste classe de variables aléatoires dont les valeurs possibles sont comprises dans l'intervalle [0, 1] (par exemple, les probabilités subjectives d'experts, relatives à un événement donné)

Suite

Définition analytique de la loi : polygone de fréquence $P(\xi = x)$ pour les variables aléatoires discrètes et densité de probabilité $f(x)$ pour les variables continues et domaine de valeurs possibles de la variable aléatoire	Valeurs approchées (estimations statistiques) des paramètres de la loi calculées sur les résultats des observations $x_1, x_2, \ldots, x_n$ de la variable aléatoire étudiée	Expressions de la moyenne $E\xi$, de la variance $V\xi$, du coefficient de dissymétrie β_1, du coefficient d'aplatissement β_2 et des covariances σ_{jk} (dans le cas multidimensionnel) en fonction des paramètres de la loi
$f(x) = \begin{cases} \dfrac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} \times \\ \times x^{a_1-1}(1- \\ -x)^{a_2-1} \text{ pour} \\ 0 \leqslant x \leqslant 1; \\ 0 \text{ pour les autres } x, \end{cases}$ $0 \leqslant x \leqslant 1$	$\hat{a}_1 = \bar{x}\left(\dfrac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$; $\hat{a}_2 = (1-\bar{x})\left(\dfrac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$, où $\bar{x}$ et s^2 sont la moyenne et la variance empiriques (méthodes moments)	$E\beta(a_1, a_2) = \dfrac{a_1}{a_1+a_2}$ $V\beta(a_1, a_2) = \dfrac{a_1 a_2}{(a_1+a_2)^2(a_1+a_2+1)}$ $\beta_1 = \dfrac{2(a_2-a_1)}{(a_1+a_2+2)} \times \dfrac{\sqrt{a_1+a_2+1}}{\sqrt{a_1 a_2}}$ $\beta_2 = \dfrac{3(a_1+a_2)}{a_1 a_2} \times \dfrac{(a_1+a_2+1)}{(a_1+a_2+2)} \times \dfrac{(a_1+1)(2a_2-a_1)}{(a_1+a_2+3)} + \dfrac{a_1(a_1-a_2)}{a_1+a_2} - 3$

CHAPITRE 7

RÉSULTATS FONDAMENTAUX DE LA THÉORIE DES PROBABILITÉS

Nous avons exposé plus haut les notions fondamentales de la théorie des probabilités y compris les lois de probabilité modèles les plus répandues dans le traitement statistique théorique et pratique des données. Le présent chapitre est consacré à la description de certaines relations entre des notions et modèles, ainsi que de certaines de leurs propriétés indispensables à la maîtrise des méthodes de modélisation probabilisto-statistique et de traitement primaire des données qui seront développées plus bas.

7.1. Inégalité de Tchébychev

Au n° 5.6.3 nous avons introduit la principale caractéristique de l'éparpillement aléatoire des valeurs d'une variable aléatoire: la variance $\sigma^2 = \mathbf{V}\xi$. De la définition de cette caractéristique il s'ensuit que la probabilité $P\{|\xi - a| \geqslant \Delta\}$, où $a = \mathbf{E}\xi$ et Δ est une quantité donnée, croît avec σ^2. Plus la variance σ^2 est élevée et plus seront probables les grands écarts des valeurs de ξ par rapport à son centre de groupement $a = \mathbf{E}\xi$. Il est évident que si l'on connaît la densité de probabilité $f_\xi(x)$ (ou le polygone de fréquence), on peut calculer exactement la probabilité d'événements de la forme $\{|\xi - a| \geqslant \Delta\}$ en l'occurrence

$$P\{|\xi - a| \geqslant \Delta\} = \begin{cases} \displaystyle\int\limits_{x:\,|x-a|\geqslant\Delta} f_\xi(x)\,\mathrm{d}x & \text{si } \xi \text{ est continue;} \\ \displaystyle\sum_{i:\,|x_i^0-a|\geqslant\Delta} p_i & \text{si } \xi \text{ est discrète.} \end{cases} \tag{7.1}$$

Si, par exemple, ξ est une *variable aléatoire normale de paramètres* (a, σ^2), la probabilité de l'événement $\{|\xi - a| \geqslant \Delta\}$ ne dépend que du rapport $\frac{\sigma}{\Delta}$.

Mais il serait intéressant d'estimer la probabilité de tels événements uniquement *à l'aide de la variance* $\sigma^2 = \mathbf{V}\xi$ *et sans connaître exactement la loi de probabilité de la variable aléatoire* ξ. La solution

de ce problème est donnée par l'inégalité de Tchébychev *)

$$P\{|\xi-a|\geqslant\Delta\}\leqslant\frac{\sigma^2}{\Delta^2}, \tag{7.2}$$

où $a = \mathrm{E}\xi$ et $\sigma^2 = \mathrm{V}\xi = \mathrm{E}(\xi - a)^2$.

La démonstration de cette inégalité est simple:

$$P\{|\xi-a|\geqslant\Delta\}=\int\limits_{x:\ |x-a|\geqslant\Delta} f_\xi(x)\,\mathrm{d}x\leqslant$$

$$\leqslant\int\limits_{x:\ |x-a|\geqslant\Delta}\left(\frac{x-a}{\Delta}\right)^2 f_\xi(x)\,\mathrm{d}x\leqslant\frac{1}{\Delta^2}\int(x-a)^2 f_\xi(x)\,\mathrm{d}x=\frac{\mathrm{V}\xi}{\Delta^2}$$

(si la variable aléatoire est discrète, on remplace les « éléments de probabilité » $f_\xi(x)\,\mathrm{d}x$ par les probabilités $p_i = P\{\xi = x_i^0\}$ et les intégrales par les sommes correspondantes).

De cette démonstration il ressort que si la *loi de probabilité de la variable aléatoire* ξ *est symétrique* (par rapport à $a = \mathrm{E}\xi$), on a les inégalités suivantes:

$$P\{\xi-a\geqslant\Delta\}=P\{a-\xi\geqslant\Delta\}\leqslant\frac{\sigma^2}{2\Delta^2}. \tag{7.2'}$$

Comme tout résultat général n'utilisant pas la forme concrète de la loi de probabilité de ξ, l'inégalité de Tchébychev *ne donne qu'une majoration grossière* des probabilités d'événements de la forme $\{|\xi - a| \geqslant \Delta\}$. Si, par exemple, l'on estime l'événement $\{|\xi - a| > 3\sigma\}$ pour la variable aléatoire normale ξ, en se servant de l'inégalité de Tchébychev, on obtient

$$P\{|\xi-a|\geqslant3\sigma\}\leqslant\frac{\sigma^2}{(3\sigma)^2}=\frac{1}{9}.$$

Il n'est pas sans intérêt de comparer ce résultat avec la *valeur exacte* de cette probabilité fournie par la table de la loi normale, soit 0,0027 : on voit donc que la valeur exacte est de 40 (!) fois inférieure à la majoration donnée par l'inégalité de Tchébychev.

7.2. Propriété de stabilité statistique des caractéristiques empiriques : loi des grands nombres et ses conséquences

On a remarqué que certaines observations (économiques, démographiques, physiques, météorologiques ou autres), bien que réalisées dans des conditions relativement homogènes, conduisaient à des résultats qui oscillaient fortement, tandis que les *moyennes d'un*

*) Cf.: Tchébychev P. *Sur les valeurs moyennes.* — Mathematitcheski sbornik, 1867, II. La même année, un article de Bienaymé a été publié dans « Journ. math. pures et appl. », XII, contenant aussi cette inégalité et l'idée de sa démonstration. C'est pourquoi cette inégalité est souvent appelée inégalité de Bienaymé-Tchébychev.

grand nombre d'observations donnaient lieu à une remarquable stabilité. A ces moyennes empiriques se rapportent toutes les caractéristiques empiriques introduites plus haut : les moments empiriques (initiaux et centrés) $\hat{m}_k(n)$ et $\hat{m}_k^{(0)}(n)$ (cf. § 5.6), la fonction de répartition empirique $\hat{F}_\xi^{(n)}(x)$, la densité de probabilité $\hat{f}_\xi^{(n)}(x)$ et les fréquences relatives $\hat{p}_i(n)$ (cf. § 5.5) (si $\hat{F}^{(n)}$, $\hat{f}^{(n)}$ et $\hat{p}_i^{(n)}$ sont interprétées comme des *moyennes* empiriques, on se rappellera seulement qu'il est possible de les exprimer par des sommes de variables aléatoires $\xi_1, \ldots, \xi_n$, où ξ_i est égale à 0 ou 1 selon que l'observation x_i tombe ou non dans le domaine des valeurs possibles défini *a priori*, cf. plus bas (7.4)).

Ce fait est traduit mathématiquement dans les diverses formes de la loi des grands nombres. La formulation de la première version particulière de cette loi est rattachée au nom du mathématicien français Poisson (Poisson S. D. *Recherches sur la probabilité de jugement en matière criminelle et en matière civile...*, Paris, Gauthier-Villars, 1837). Dans la formulation qui est proposée plus bas, ce théorème a été démontré pour la première fois par A. Hintchin (cf. Hintchin A. *Sur la loi des grands nombres.* Comptes rendus de l'Académie des Sciences, **189** (1929), 477-479).

7.2.1. Loi des grands nombres. Soient $\xi_1, \xi_2, \ldots, \xi_n$ une suite de variables aléatoires équiréparties. Si la moyenne $a = \mathbf{E}\xi_i$ existe, la moyenne arithmétique $\frac{\xi_1+\cdots+\xi_n}{n}$ *converge en probabilité* (on dit encore stochastiquement) vers a lorsque $n \to \infty$, autrement dit, pour tous $\varepsilon > 0$ et $\delta > 0$ aussi petits que l'on veut, on peut exhiber un indice n_0 à partir duquel (c'est-à-dire pour tout $n \geqslant n_0$) on aura

$$P\left\{\left|\frac{\xi_1+\cdots+\xi_n}{n} - a\right| < \varepsilon\right\} > 1-\delta. \tag{7.3}$$

La démonstration de cette proposition n'apporte aucune complication si l'on exige en plus que les variables aléatoires ξ_i admettent une variance finie, c'est-à-dire $\mathbf{V}\xi_i = \sigma^2 < \infty$. En effet, pour prouver (7.3) dans ce cas, il suffit d'appliquer l'inégalité de Tchébychev (7.2) à la variable aléatoire $\overline{\xi}(n) = (\xi_1 + \ldots + \xi_n)/n$. Il est immédiat que $\mathbf{E}\overline{\xi}(n) = a$ et $\mathbf{V}\overline{\xi}(n) = \sigma^2/n$, donc, en vertu de (7.2),

$$P\{|\overline{\xi}(n) - a| \geqslant \varepsilon\} \leqslant \frac{\mathbf{V}\overline{\xi}(n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

En prenant $n_0 \geqslant \frac{\sigma^2}{\delta\varepsilon^2}$, on voit aussitôt que (7.3) est réalisée pour tous ε et δ.

La démonstration de (7.3) dans le cas général est accessible, par exemple, dans [54].

A titre de conséquence de la loi des grands nombres, considérons l'important résultat suivant qui explique l'effet de stabilité des *fréquences relatives* *).

7.2.2. Théorème de J. Bernoulli. Supposons qu'on effectue n expériences aléatoires (ou observations d'une variable aléatoire ξ) dont l'issue est la réalisation ou non d'un événement A (par exemple, l'événement qui consiste en ce que $\xi \in \Delta X$, où ΔX est un domaine mesurable donné des valeurs possibles de la variable aléatoire ξ). Lorsque le nombre d'expériences n tend vers l'infini, la fréquence relative $\hat{p}^{(n)}(A)$ d'apparition de l'événement A converge stochastiquement vers la probabilité $p(A)$ de cet événement, c'est-à-dire que pour tous nombres $\varepsilon > 0$ et $\delta > 0$ aussi petits que l'on veut, on peut exhiber un indice n_0 tel que pour tout $n \geqslant n_0$, l'on a

$$P\{|\hat{p}^{(n)}(A) - p(A)| < \varepsilon\} > 1 - \delta. \tag{7.3'}$$

La démonstration de cette proposition résulte de (7.3) si pour variables aléatoires on prend des variables ξ_i telles que

$$\xi_i = \begin{cases} 1 \text{ si l'événement } A \text{ se produit à l'issue de la } i\text{-ième} \\ \quad \text{expérience;} \\ 0 \text{ sinon.} \end{cases} \tag{7.4}$$

De la définition il s'ensuit que toutes ces variables aléatoires $\xi_1, \ldots, \xi_n$ suivent la même loi de probabilité, et, en particulièr,

$$\begin{aligned} \mathbf{E}\xi_i &= 1 \cdot p(A) + 0 \cdot (1 - p(A)) = p(A); \\ \mathbf{V}\xi_i &= (1 - p(A))^2 \cdot p(A) + (0 - p(A))^2 (1 - p(A)) = \\ &= p(A)(1 - p(A)). \end{aligned} \tag{7.5}$$

Il est évident que dans ce cas $\overline{\xi}(n) = (\xi_1 + \ldots + \xi_n)/n$ n'est autre que la fréquence relative $\hat{p}^{(n)}(A)$ d'apparition de l'événement A à l'issue de n expériences aléatoires, et, de plus,

$$\begin{aligned} \mathbf{E}\overline{\xi}(n) &= \mathbf{E}\hat{p}^{(n)}(A) = p(A); \\ \mathbf{V}\overline{\xi}(n) &= \mathbf{V}\hat{p}^{(n)}(A) = \frac{p(A)(1 - p(A))}{n}. \end{aligned} \tag{7.6}$$

En appliquant la loi des grands nombres (7.3) aux variables aléatoires (7.4), on obtient, compte tenu de (7.5) et (7.6), la démonstration du théorème de Bernoulli (7.3').

7.2.3. Stabilité statistique des caractéristiques empiriques. La loi des grands nombres et le théorème de Bernoulli permettent de

*) Le théorème de Bernoulli est apparu bien avant le théorème fort (7.3), plus exactement, en 1713. Mais dans les traités modernes il est plus commode de le présenter comme un simple corollaire du théorème (7.3).

justifier théoriquement la stabilité des principales caractéristiques empiriques d'une loi : la moyenne, la variance, le coefficient de dissymétrie, le coefficient d'aplatissement, la fonction de répartition et la densité de probabilité, construits sur l'échantillon $x_1, x_2, \ldots, x_n$. Ceci étant, comme toujours, dès qu'il sera question d'étudier des caractéristiques empiriques, primo, nous admettrons que nous avons affaire à un échantillon composé d'observations *indépendantes*, secundo, nous interpréterons l'échantillon au sens *hypothétique*, c'est-à-dire comme un ensemble d'observations indépendantes qui *auraient pu être réalisées* sur la variable aléatoire analysée (cf. note du n° 5.6.4). Dans cette interprétation, les observations $x_1, x_2, \ldots, x_n$ sont des variables aléatoires indépendantes équiréparties justiciables des relations (7.3) et (7.3'). Montrons comment appliquer la loi des grands nombres et le théorème de Bernoulli à l'établissement de la stabilité statistique des principales caractéristiques empiriques.

a. *Stabilité des moments initiaux empiriques* $\hat{m}_k(n)$ *et de toutes fonctions rationnelles de ces moments.* Supposons qu'existent tous les moments $m_k = \mathbf{E}\xi^k$ $(k = 1, 2, \ldots, 2k_0)$ d'une variable aléatoire ξ pour k_0 donné. En appliquant alors la loi des grands nombres aux variables aléatoires $\xi_1 = x_1^k$, $\xi_2 = x_2^k, \ldots, \xi_n = x_n^k$, où x_i est le résultat de la i-ième observation de la variable aléatoire ξ, on obtient immédiatement la démonstration de la convergence stochastique de tous les moments initiaux empiriques $\hat{m}_k(n) = (\sum_{i=1}^{n} x_i^k)/n$ vers les moments théoriques respectifs $m_k = \mathbf{E}\xi^k$ $(k = 1, 2, \ldots, k_0)$.

L'application directe de la loi des grands nombres aux observations *centrées* $x_1 - \bar{x}(n), \ldots, x_n - \bar{x}(n)$ est impossible, car ces observations deviennent *dépendantes* une fois centrées.

Mais en appliquant le théorème de E. Slutsky *) qui dit que la convergence stochastique $(n \to \infty)$ de variables aléatoires $\xi_i(n)$ vers des nombres constants a_i $(i = 1, 2, \ldots, k_0)$ entraîne la convergence stochastique de toute fonction rationnelle $\varphi(\xi_1(n), \xi_2(n), \ldots, \xi_{k_0}(n))$ vers sa valeur $\varphi(a_1, a_2, \ldots, a_{k_0})$ au point $(a_1, a_2, \ldots, a_{k_0})$ (si celle-ci existe), on obtient immédiatement la démonstration de la convergence stochastique de tous les moments centrés empiriques, du coefficient de dissymétrie et du coefficient d'aplatissement vers les valeurs théoriques respectives (si elles existent). Ceci étant, on tient évidemment compte du fait que les moments centrés, le coefficient de dissymétrie et le coefficient d'aplatissement sont des fonctions rationnelles des moments initiaux (cf. relations (5.22)).

---

*) Slutsky E. *Über stochastische Asymptoten und Grenzwerte.* — Metron, 5, N° 3 (1925), 3.

b. *La stabilité de la fonction de répartition, de la densité de probabilité et des fréquences relatives empiriques*, c'est-à-dire leur convergence (lorsque la taille de l'échantillon sur lequel elles ont été construites tend vers l'infini) vers les fonctions théoriques et les probabilités respectives résulte directement de (7.3) et (7.3'). Prouvons-le sur l'exemple de la fonction de répartition empirique $\hat{F}^{(n)}(x)$. Introduisons les variables aléatoires (7.4), où l'événement A est défini comme $\{x_i < x\}$, c'est-à-dire que

$$\xi_i = \xi(x_i) = \begin{cases} 1 & \text{si } x_i < x; \\ 0 & \text{si } x_i \geqslant x. \end{cases}$$

Il est évident que $\xi_1, \xi_2, \ldots, \xi_n$ sont des variables aléatoires équiréparties indépendantes et, de plus, $\mathbf{E}\xi_i = F(x)$ et $\mathbf{V}\xi_i = F(x)(1 - F(x))$, où $F(x) = P\{\xi < x\}$ est la fonction de répartition de la variable aléatoire *étudiée* ξ. Il est aisé de voir que $\hat{F}^{(n)}(x) = (\sum_{i=1}^{n} \xi_i)/n = \bar{\xi}(n)$ et, par suite, la loi des grands nombres nous dit que $\hat{F}^{(n)}(x) \to F(x)$ stochastiquement lorsque $n \to \infty$.

7.3. Rôle particulier de la loi normale : théorème limite central

Les résultats du § 7.2 s'interprètent *grosso modo* comme suit : lorsqu'on prend la moyenne d'un grand nombre n de variables aléatoires, l'éparpillement incontrôlable, qui est caractéristique aux valeurs prises par ces variables, se fait de moins en moins sentir et *à la limite* lorsque $n \to \infty$, il disparaît complètement ou, comme on a encore coutume de dire, la *variable aléatoire dégénère en variable non aléatoire*. Mais cette dispersion aléatoire autour de la moyenne subsiste lorsque le nombre n de variables est *fini*. Ce qui pose le problème de son étude (asymptotique lorsque $n \to \infty$). Un résultat fondamental (connu sous le nom de « théorème limite central ») a été publié pour la première fois dans l'ouvrage déjà cité de Laplace (1812). Ce résultat dit que pour une vaste classe de variables aléatoires indépendantes $\xi_1, \ldots, \xi_n$ la loi de probabilité limite (lorsque $n \to \infty$) de leur somme normalisée tend vers la loi normale *quelle que soit la loi de probabilité* des termes de cette somme. Mais cette formulation appelle quelques précisions : que veut dire somme « normalisée » de variables aléatoires et *dans quel sens* la loi de probabilité d'une variable aléatoire tend-elle vers la loi d'une autre ? Il existe plusieurs énoncés rigoureux du théorème limite central qui se distinguent par

le degré de généralité et par la forme des contraintes adoptées. Nous citons ici l'énoncé de Lindeberg et Lévy *).

7.3.1. Théorème limite central. Si $\xi_1, \xi_2, \ldots, \xi_n$ sont des variables aléatoires indépendantes équidistribuées de moyenne $\mathbf{E}\xi_i = a$ et de variance $\mathbf{V}\xi_i = \sigma^2$, alors la fonction de répartition de la variable aléatoire

$$\bar{\xi}^*(n) = \frac{\frac{1}{n}(\xi_1 + \ldots + \xi_n) - a}{\sigma/\sqrt{n}} = \frac{(\bar{\xi}(n) - a)\sqrt{n}}{\sigma}$$

tend pour $n \to \infty$ vers la fonction de répartition de la loi normale pour toute valeur donnée de leurs arguments, c'est-à-dire que

$$F_{\bar{\xi}^*(n)}(x) \to \Phi(x) \text{ lorsque } n \to \infty$$

pour toute valeur de x, où

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Ainsi, le théorème limite central donne une description mathématique rigoureuse des conditions induisant le mécanisme de la loi normale (cf. la discussion non formelle de ces conditions au n° 6.1.5). Ce théorème légitime notamment le rôle central tenu par la loi normale dans la théorie et la pratique des recherches statistiques. Le théorème limite central peut être considéré (après la loi des grands nombres) comme une précision du comportement stochastique de la moyenne arithmétique d'une série de variables aléatoires.

Le théorème limite central peut être développé dans plusieurs directions : au cas où les valeurs aléatoires ne sont pas équiréparties (formulation de Liapounov) ; au cas où les composantes ξ_i ne sont pas indépendantes ; au cas, enfin, où les variables aléatoires ξ_i sont multidimensionnelles.

7.3.2. Théorème limite central multidimensionnel (cf. [16]). Soient $\xi_1, \xi_2, \ldots, \xi_n$ des variables aléatoires indépendantes équiréparties p-dimensionnelles de vecteur des moyennes $M = \mathbf{E}\xi_i = (\mathbf{E}\xi_i^{(1)}, \mathbf{E}\xi_i^{(2)}, \ldots, \mathbf{E}\xi_i^{(p)})'$ et de matrice des covariances $\Sigma = \mathbf{E}\{(\xi_i - M)(\xi_i - M)'\}$. Sous ces conditions, la fonction de répartition conjointe du vecteur aléatoire $\xi^*(n) = [\sum_{i=1}^{n}(\xi_i - M)]/\sqrt{n}$ converge lorsque $n \to \infty$ (pour toute valeur de X) vers la fonction de

*) Lévy Paul. *Calcul des probabilités*. Paris, 1925 ; Lindeberg J.W. *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechung*. — Math. Zeitschr., 15 (1922), 211.

répartition conjointe d'une variable aléatoire de vecteur des moyennes $\mathbf{0} = (0, 0, \ldots, 0)'$ et de matrice des covariances Σ.

Remarque 1. Le théorème limite central doit être manipulé avec précaution dans les études statistiques.

Primo, si la forme *limite* de la loi de probabilité de la somme de variables aléatoires est sous certaines conditions toujours normale et ne dépend pas de la loi des termes de cette somme, la vitesse de convergence de la loi de cette somme vers la loi normale dépend, elle, essentiellement de la loi des termes de cette somme. Si, par exemple, des variables aléatoires sont équiréparties, la somme de 6 à 10 d'entre elles suit approximativement une loi normale, alors qu'il en faudrait plus de 100 si elles suivaient une loi du χ^2.

Secundo, le théorème limite central n'est généralement pas recommandé pour approcher les probabilités aux ailes des lois, c'est-à-dire dans l'estimation des probabilités d'événements de la forme $\{\bar{\xi}^*(n) < \tilde{x}_{\min}\}$ et $\{\bar{\xi}^*(n) \geqslant \tilde{x}_{\max}\}$, où $\tilde{x}_{\min}$ et $\tilde{x}_{\max}$ sont des valeurs possibles voisines respectivement des bornes gauche et droite de l'intervalle de variation de la variable aléatoire $\bar{\xi}^*(n)$. Vu que dans ce cas les valeurs des probabilités $P\{\bar{\xi}^*(n) < \tilde{x}_{\min}\} = F_{\bar{\xi}^*(n)}(\tilde{x}_{\min})$ et $P\{\bar{\xi}^*(n) > \tilde{x}_{\max}\} = 1 - F_{\bar{\xi}^*(n)}(\tilde{x}_{\max})$ sont petites, la petitesse des différences $F_{\bar{\xi}^*(n)}(\tilde{x}_{\min}) - \Phi(\tilde{x}_{\min})$ et $F_{\bar{\xi}^*(n)}(\tilde{x}_{\max}) - \Phi(\tilde{x}_{\max})$ (qui résulte du théorème limite central) n'entraîne aucunement celle des erreurs relatives d'approximation

$$\frac{F_{\bar{\xi}^*(n)}(\tilde{x}_{\min})}{\Phi(\tilde{x}_{\min})} \quad \text{et} \quad \frac{1 - F_{\bar{\xi}^*(n)}(\tilde{x}_{\max})}{1 - \Phi(\tilde{x}_{\max})},$$

qui sont généralement excessivement grandes. Supposons, par exemple, que $\bar{\xi}^*(n)$ est le revenu moyen normalisé par personne dans une famille ($\xi_1, \xi_2, \ldots$ sont respectivement le salaire des membres actifs de la famille et les autres composantes du revenu familial) et étudions le taux q des familles ayant un revenu très élevé, plus exactement, un revenu supérieur à un niveau assez élevé $\tilde{x}_{\max}$. Une analyse montre que la valeur exacte de q est $q = 1 - F_{\bar{\xi}^*(n)}(\tilde{x}_{\max}) = 0{,}03$, alors que l'approximation normale respective donne $\hat{q} = 1 - \Phi(\tilde{x}_{\max}) = 0{,}003$. La différence $q - \hat{q}$ est petite en soi (ainsi qu'il ressort du théorème limite central), mais l'erreur relative de l'approximation normale est de 1000 % ! Cet avertissement vaut surtout quand on essaye d'appliquer les approximations normales aux calculs des dépendances de type « résistance limite (ou capacité

d'accueil) d'un système — probabilité de destruction (ou de refus d'une demande) ».

Remarque 2. Le théorème limite central permet d'étudier les relations asymptotiques existant entre les divers modèles de lois de probabilité (cf. chapitre 6), d'une part, et la loi normale, de l'autre. En se prévalant du théorème limite central, on peut, en particulier, expliquer les faits suivants qui sont d'une grande utilité pratique.

1. La loi d'une variable aléatoire binomiale $\xi(p, n)$ de paramètres (p, n) est asymptotiquement normale *) (par rapport à $n \to \infty$) de paramètres $\mathbf{E}\xi(p, n) = np$ et $\mathbf{V}\xi(p, n) = np(1-p)$. Ce résultat est connu sous le nom de *théorème de Moivre-Laplace* (la démonstration a été donnée par Moivre en 1733, alors que le théorème limite central n'était pas encore connu) et découle directement du théorème limite central appliqué à la variable aléatoire (7.4) compte tenu de (7.5).

2. La loi d'une variable aléatoire poissonnienne $\xi(\lambda)$ est asymptotiquement normale (par rapport à $\lambda \to \infty$) de paramètres $\mathbf{E}\xi(\lambda) = \lambda$ et $\mathbf{V}\xi(\lambda) = \lambda$.

3. La loi d'une variable aléatoire hypergéométrique $\xi(N, M, n)$ de paramètres (N, M, n) est asymptotiquement normale (par rapport à $N \to \infty$, $M \to \infty$, $\frac{M}{N} \to p > 0$ et $n \to \infty$) de paramètres $\mathbf{E}\xi(N, M, n) = np$ et $\mathbf{V}\xi(N, M, n) = np(1-p)$.

4. La fonction de répartition d'une variable aléatoire polynomiale normalisée k-dimensionnelle de paramètres $(p_1, p_2, \ldots, p_k; n)$

$$\xi = \left(\frac{\xi^{(1)} - np_1}{\sqrt{np_1}}, \ldots, \frac{\xi^{(k)} - np_k}{\sqrt{np_k}}\right)$$

tend, lorsque $n \to \infty$, vers la fonction de répartition d'une loi normale impropre (dégénérée) k-dimensionnelle dont les valeurs moyennes sont nulles et la matrice des covariances

$$\Sigma = \mathbf{I}_k - \begin{pmatrix} \sqrt{p_1} \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{p_k} \end{pmatrix} (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k})$$

possède un rang égal à $k-1$ (cf., par exemple, [16], « Application » **).

---

*) On dit qu'une variable aléatoire $\xi(n)$ dépendant d'un paramètre n est *asymptotiquement normale (par rapport à n)* s'il existe des variables non aléatoires de normalisation $A(n)$ et $B(n)$, telles que la fonction de répartition de la variable aléatoire $\eta(n) = A(n) \cdot \xi(n) + B(n)$ tend, lorsque $n \to \infty$, vers la fonction de répartition de la loi normale réduite pour toute valeur donnée de leur argument x.

**) Ici et dans la suite le symbole $\mathbf{I}_k$ désignera la matrice unité de dimension k.

5. La loi d'une variable aléatoire $\chi^2(m)$ est asymptotiquement normale (par rapport à $m \to \infty$) de paramètres $\mathbf{E}\chi^2(m) = m$ et $\mathbf{V}\chi^2(m) = 2m$.

6. La loi d'une variable aléatoire $t(m)$ est asymptotiquement normale (par rapport à $m \to \infty$) de paramètres $\mathbf{E}t(m) = 0$ et $\mathbf{V}t(m) = 1$.

7.4. Loi de probabilité de variables aléatoires qui sont fonctions de variables aléatoires connues

Dans les recherches statistiques théoriques et pratiques, il est très important de savoir calculer la loi de probabilité de fonctions dépendant de variables aléatoires dont on connaît la répartition. Ceci constitue l'essence de la théorie de l'estimation statistique et de la vérification des hypothèses statistiques (cf. chap. 8), puisque l'*estimation statistique* et la *statistique critique*, utilisées respectivement pour l'estimation des valeurs inconnues des paramètres et pour la construction de tests d'hypothèses statistiques, sont des *fonctions des observations* de la variable aléatoire ξ étudiée. Pour en tirer le meilleur parti et connaître leurs propriétés statistiques, nous devons être en mesure de reconstituer leur loi de probabilité d'après celle de la variable aléatoire ξ (donc, d'après ses observations). On décrit plus bas les principales règles qui nous guideront dans la résolution de ce problème.

1. Supposons qu'une variable aléatoire η est une *fonction continue monotone strictement croissante* d'une variable aléatoire ξ donnée dont la fonction de répartition $F_\xi(x)$ est partout différentiable, c'est-à-dire que $\eta = g(\xi)$. A toute valeur possible x de la variable aléatoire ξ sera associée une valeur possible $y = g(x)$ de η.

La fonction $\eta = g(\xi)$ étant continue et monotone, on peut déterminer de façon unique ξ en fonction de η à l'aide de la transformation réciproque g^{-1} de g, soit $\xi = g^{-1}(\eta)$. On a une relation analogue entre les valeurs possibles de ces variables, soit $x = g^{-1}(y)$.

Essayons d'exprimer la fonction de répartition $F_\eta(y)$ en termes de fonctions $F_\xi(x)$, g et g^{-1}:

$$F_\eta(y) = P\{\eta < y\} = P\{g^{-1}(\eta) < g^{-1}(y)\} = \\ = P\{\xi < g^{-1}(y)\} = F_\xi(g^{-1}(y)). \qquad (7.7)$$

La dérivation des deux membres de (7.7) par rapport à y nous donne

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy}. \qquad (7.8)$$

Les mêmes raisonnements appliqués à une *fonction monotone strictement décroissante* $g(x)$ nous conduisent à la modification sui-

vante de la formule (7.8):

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \left(-\frac{dg^{-1}(y)}{dy}\right). \tag{7.8'}$$

On peut regrouper les formules (7.8) et (7.8') en une seule valable pour toute application bijective g:

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \left|\frac{dg^{-1}(y)}{dy}\right|. \tag{7.8''}$$

Exemple 7.1. Calculer la densité de probabilité d'une variable aléatoire $\eta = e^\xi$ sachant que ξ est une variable aléatoire normale de paramètres (a, σ^2).

Dans cet exemple $g(x) = e^x$, $g^{-1}(y) = \ln y$, donc

$$\frac{dg^{-1}(y)}{dy} = \frac{d(\ln y)}{dy} = \frac{1}{y}.$$

En portant ceci dans (7.8''), on trouve

$$f_\eta(y) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{\frac{-(\ln y - a)^2}{2\sigma^2}} \cdot \frac{1}{y},$$

c'est-à-dire la densité de la *loi lognormale* (cf. n° 6.1.6).

Exemple 7.2. Calculer la densité de probabilité de la variable aléatoire $\eta = a + b\xi$ sachant la densité $f_\xi(x)$ de la variable aléatoire ξ.

Ici $g(x) = a + bx$, $g^{-1}(y) = (y - a)/b$ et $(dg^{-1}(y)/dy) = 1/b$. Les relations (7.7) et (7.8'') nous donnent

$$\begin{aligned} F_\eta(y) &= F_\xi\left(\frac{y-a}{b}\right); \\ f_\eta(y) &= f_\xi\left(\frac{y-a}{b}\right) \cdot \left|\frac{1}{b}\right|. \end{aligned} \tag{7.9}$$

Cette règle de calcul des fonctions de répartition et des densités de probabilité permet, en particulier, de se servir des tables de la loi normale *réduite* pour déterminer les valeurs prises par la fonction de répartition et la densité de probabilité d'une variable aléatoire normale $\xi(a, \sigma^2)$ de paramètres *arbitraires* (a, σ^2). On remarquera que ξ figure la variable aléatoire normale réduite $\xi(0, 1)$ et η, la variable aléatoire normale $\xi(a, \sigma^2)$, c'est-à-dire que

$$\xi(a, \sigma^2) = a + \sigma \cdot \xi(0, 1). \tag{7.10}$$

2. *Si l'application $\eta = g(\xi)$ n'est pas bijective, l'établissement de formules générales ne présente aucun intérêt.* Les problèmes doivent être examinés alors cas par cas. Considérons, par exemple, le cas $\eta = \xi^2$:

$$F_\eta(y) = P\{\eta < y\} = P\{|\xi| < \sqrt{y}\} = P\{-\sqrt{y} < \xi < \sqrt{y}\} = F_\xi(\sqrt{y}) - F_\xi(-\sqrt{y}).$$

Donc,

$$f_\eta(y) = f_\xi(\sqrt{y})\cdot\frac{1}{2\sqrt{y}} + f_\xi(-\sqrt{y})\cdot\frac{1}{2\sqrt{y}} =$$
$$= \frac{1}{2\sqrt{y}}(f_\xi(\sqrt{y}) + f_\xi(-\sqrt{y})).$$

L'application de cette formule à une variable aléatoire normale réduite ξ nous donne

$$f_\eta(y) = \frac{1}{2\sqrt{y}}\cdot 2\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}} = \frac{\left(\frac{1}{2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)}y^{1/2-1}e^{-1/2y},$$

densité qui est un cas particulier de celle de la loi gamma de paramètres $a = b = \frac{1}{2}$ (cf. 6.2.5).

3. Généralisons la formule (7.8″). Soit $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ une variable aléatoire p-dimensionnelle dont on connaît la fonction de répartition $F_\xi(X)$ et la densité de probabilité $f_\xi(X)$, et supposons qu'une autre variable aléatoire p-dimensionnelle $\eta = (\eta^{(1)}, \ldots, \eta^{(p)})$ est une fonction *vectorielle* continue donnée $g(\xi)$, c'est-à-dire que

$$\eta^{(1)} = g_1(\xi^{(1)}, \ldots, \xi^{(p)});$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\eta^{(p)} = g_p(\xi^{(1)}, \ldots, \xi^{(p)}).$$

On admet que l'application $\eta = g(\xi)$ est bijective, c'est-à-dire qu'existe l'application réciproque g^{-1} qui permet de remonter de η à ξ:

$$\xi^{(1)} = g_1^{-1}(\eta^{(1)}, \ldots, \eta^{(p)});$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\xi^{(p)} = g_p^{-1}(\eta^{(1)}, \ldots, \eta^{(p)}).$$

Respectivement, entre les « valeurs » possibles multidimensionnelles $X = (x^{(1)}, \ldots, x^{(p)})$ et $Y = (y^{(1)}, \ldots, y^{(p)})$ des variables aléatoires ξ et η on a les relations vectorielles

$$Y = g(X) \text{ et } X = g^{-1}(Y).$$

La densité de probabilité conjointe des variables aléatoires $\eta = (\eta^{(1)}, \ldots, \eta^{(p)})$ est alors égale à

$$f_\eta(y^{(1)}, \ldots, y^{(p)}) = f_\xi(g_1^{-1}(Y), g_2^{-1}(Y), \ldots, g_p^{-1}(Y))\cdot|J(Y)|, \quad (7.11)$$

où $|J(Y)|$ est la valeur absolue du *jacobien*:

$$J(Y)=\begin{vmatrix} \frac{\partial g_1^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_1^{-1}(Y)}{\partial y^{(2)}} & \cdots & \frac{\partial g_1^{-1}(Y)}{\partial y^{(p)}} \\ \frac{\partial g_2^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_2^{-1}(Y)}{\partial y^{(2)}} & \cdots & \frac{\partial g_2^{-1}(Y)}{\partial y^{(p)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_p^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_p^{-1}(Y)}{\partial y^{(2)}} & \cdots & \frac{\partial g_p^{-1}(Y)}{\partial y^{(p)}} \end{vmatrix}.$$

4. *Etablissons la loi de la somme de deux variables aléatoires indépendantes (formule de composition).* Soient ξ_1 et ξ_2 des variables aléatoires indépendantes de densités de probabilité respectives $f_{\xi_1}(x)$ et $f_{\xi_2}(y)$. On demande la composition de ces densités, c'est-à-dire la densité de probabilité de la variable aléatoire $\eta = \xi_1 + \xi_2$. Nous devons en fait considérer la densité de probabilité conjointe $f_{\xi_1, \xi_2}(x, y)$ et, pour déterminer la fonction de répartition de η, trouver dans le plan xOy le domaine de valeurs possibles de (ξ_1, ξ_2) correspondant à l'événement $\{\eta < z\}$. Ce domaine est hachuré et désigné par A_z sur la figure 7.1. On obtient

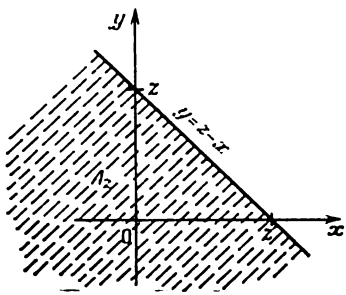


Fig. 7.1. Tout point du domaine A_z du plan xOy correspond à l'événement $\{\xi_1 + \xi_2 < z\}$

$$F_\eta(z) = P\{\eta < z\} = P\{\xi_1 + \xi_2 < z\} = P\{(\xi_1, \xi_2) \in A_z\} =$$
$$= \iint\limits_{A_z} f_{(\xi_1, \xi_2)}(x, y)\,dx\,dy = \iint\limits_{A_z} f_{\xi_1}(x) \cdot f_{\xi_2}(y)\,dx\,dy =$$
$$= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{z-x} f_{\xi_1}(x) f_{\xi_2}(y)\,dx\,dy = \int\limits_{-\infty}^{\infty} f_{\xi_1}(x) \left(\int\limits_{-\infty}^{z-x} f_{\xi_2}(y)\,dy \right) dx. \quad (7.12)$$

On s'est servi de l'identité $f_{(\xi_1, \xi_2)}(x, y) = f_{\xi_1}(x) f_{\xi_2}(y)$ (qui est valable, car ξ_1 et ξ_2 sont indépendantes).

La dérivation par rapport à z des deux membres de (7.21) nous donne

$$f_\eta(z) = \int\limits_{-\infty}^{\infty} f_{\xi_1}(x) f_{\xi_2}(z-x)\,dx. \quad (7.13)$$

La formule (7.13) s'appelle formule de composition de deux répartitions ou *formule de convolution*. Le produit de convolution des lois de probabilité est souvent désigné par

$$f_\eta = f_{\xi_1} * f_{\xi_2}.$$

La formule (7.13) nous permet d'établir la propriété d'« autoreproduction » des lois de Gauss et de Cauchy mentionnée aux n°s 6.1.5 et 6.1.10, savoir que la somme de variables aléatoires normales est normale et la somme de variables aléatoires de Cauchy équiréparties est une variable de Cauchy, ainsi que les formules de la densité de probabilité des sommes de variables aléatoires uniformes citées au n° 6.1.7.

Conclusions

1. On distinguera trois types de résultats fondamentaux en théorie des probabilités:

les résultats *préasymptotiques* qui permettent d'analyser les principales lois régissant la variable aléatoire à travers ses caractéristiques numériques fondamentales — moyenne, variance, etc. — sans en appeler à la loi de probabilité;

les résultats *asymptotiques* grâce auxquels il est possible d'analyser les lois fondamentales de la somme d'un grand nombre de variables aléatoires (en l'occurrence, d'établir leur stabilité asymptotique, c'est-à-dire leur convergence vers une valeur constante lorsque le nombre des termes de cette somme croît, ou de décrire la forme asymptotique de la loi de probabilité de ces sommes) sans connaître exactement la loi de probabilité des termes de cette somme;

les résultats liés aux *transformations des variables aléatoires* qui nous aident à trouver la loi de probabilité de fonctions dépendant de variables aléatoires dont on connaît la répartition conjointe.

2. Les résultats du premier type sont représentés dans cet ouvrage par l'*inégalité de Tchébychev* qui permet d'estimer la probabilité que l'écart entre la valeur prise par la variable aléatoire ξ et sa moyenne $a = \mathrm{E}\xi$ soit supérieur à une quantité donnée Δ, c'est-à-dire $P\{|\xi - a| \geqslant \Delta\}$, sans connaître la loi de probabilité de ξ (cf. (7.2)).

3. La *loi des grands nombres* et ses conséquences se rapportent au premier niveau (le moins profond) des résultats asymptotiques et permettent d'établir la convergence stochastique de sommes normalisées d'un grand nombre de variables aléatoires vers des valeurs constantes — lorsque le nombre de termes croît — pratiquement indépendamment de la loi de probabilité de ces variables. De ces résultats on déduit directement, en particulier, l'importante propriété de stabilité statistique des principales caractéristiques numériques empiriques de la variable aléatoire étudiée: moyenne, variance, fréquences relatives, etc. (cf. (7.3), (7.3')).

4. Le *théorème limite central* fait partie du niveau suivant (plus profond) des résultats asymptotiques. Il affirme, en particulier, que la loi de probabilité de la somme normalisée d'un grand nombre de variables aléatoires (lorsque ce nombre croît) tend vers la loi

normale pratiquement indépendamment de la répartition de ces variables (cf. n° 7.3.1).

Mais le théorème limite central doit être manipulé avec précaution dans la pratique: premièrement, on peut souvent choisir des approximations relativement simples (et plus précises que les normales !) de la loi de probabilité d'une somme d'un nombre *fini peu élevé* de variables aléatoires; deuxièmement, le théorème limite central *fonctionne mal aux ailes des lois*, c'est-à-dire lorsqu'on estime les probabilités des grands écarts de la somme des variables aléatoires par rapport à sa moyenne.

5. Le principal résultat de la *théorie des transformations des variables aléatoires* est la règle (7.8″) (ou (7.11) dans le cas multidimensionnel) qui permet de calculer la loi de probabilité (la densité de probabilité ou le polygone de fréquence) d'une variable aléatoire qui est une fonction donnée d'un ensemble de variables aléatoires dont on connaît la loi de probabilité conjointe.

TROISIÈME PARTIE

ÉLÉMENTS DE STATISTIQUE MATHÉMATIQUE

CHAPITRE 8

ESTIMATION STATISTIQUE DES PARAMÈTRES

L'un des principaux objectifs que se fixe le chercheur en abordant le traitement statistique des données initiales est de décrire avec *laconisme* les propriétés de la population (ou du phénomène) étudiée, c'est-à-dire de représenter l'ensemble des données à traiter sous la forme d'un *dossier de caractéristiques* construites sur la base de ces données initiales. Ceci étant, il est souhaitable que la perte d'information soit minimale. Les caractéristiques mentionnées sont des fonctions des observations initiales $X_1, X_2, \ldots, X_n$ et s'appellent *statistiques* (ce terme est donc employé dans trois acceptions : discipline scientifique, information initiale et fonction des observations). Au chapitre précédent nous avons eu affaire aux statistiques : les caractéristiques *empiriques* de la population générale, c'est-à-dire les moyennes, la variance, les coefficients de dissymétrie et d'aplatissement, la covariance et la corrélation et, enfin, la fonction de répartition et la densité de probabilité empiriques (cf. formules (5.6), (5.8), (5.19), (5.20′), (5.21′), (5.23′), (5.34′), (5.35′), (5.36′), (5.37′), (5.38′)).

Pour décrire laconiquement l'information contenue dans le fichier, nous disposons de tout un *arsenal de méthodes appliquées* de statistique mathématique : choix et justification du modèle mathématique du mécanisme du phénomène étudié (cf. chap. 3) ; étude des propriétés du système analysé ou du mécanisme de fonctionnement par simulation sur ordinateur (§§ 3.2 et 6.3) ; visualisation des données initiales afin d'énoncer les hypothèses de travail sur le mécanisme du phénomène étudié (cf. § 10.5) et autres méthodes de statistique descriptive (étude des lois empiriques, manipulation de variables de types divers, etc., cf. §§ 10.2, 10.3) ; analyse de la nature des données traitées (chap. 11) ; description des relations statistiques liant les variables analysées, etc. Toutes ces méthodes relèvent d'une manière ou d'une autre de *deux domaines fondamentaux de la statistique mathématique* : 1) de la *théorie de l'estimation statisti-*

que des valeurs inconnues des paramètres participant à la description du modèle analysé ; 2) de la *théorie des tests des hypothèses statistiques concernant les paramètres ou la nature du modèle analytique.*

Le présent chapitre est consacré à l'exposé des principaux éléments du premier domaine mentionné.

8.1. Notions préliminaires d'estimation statistique des paramètres

8.1.1. Position du problème. Soit donné un échantillon

$$X_1, X_2, \ldots, X_n \tag{8.1}$$

et supposons que les caractères de la population générale peuvent être décrits à l'aide des équations (le *modèle mathématique*)

$$\mathbf{M}(X, \Theta) = 0, \tag{8.2}$$

où $X_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)})'$ est la i-ième observation de l'échantillon (8.1), X la valeur courante (c'est-à-dire substituée par nous) de la variable aléatoire p-dimensionnelle étudiée, $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ un paramètre k-dimensionnel participant dans l'écriture du modèle (8.2) et dont les valeurs sont inconnues avant le tirage de l'échantillon (8.1).

Le problème d'estimation statistique des paramètres inconnus Θ au vu de l'échantillon (8.1) consiste *grosso modo* à construire une fonction vectorielle k-dimensionnelle $\hat{\Theta}(X_1, \ldots, X_n) = (\hat{\theta}^{(1)}(X_1, \ldots, X_n), \ldots, \hat{\theta}^{(k)}(X_1, \ldots, X_n))'$ des observations (8.1) qui *dans un certain sens donne les meilleures valeurs approchées* des véritables valeurs (inconnues) des paramètres $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})'$. Pour l'instant, on ne précise pas dans quel sens les valeurs approchées $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(k)}$ respectivement des paramètres $\theta^{(1)}, \ldots, \theta^{(k)}$ sont les meilleures.

Le modèle (8.2) peut être un modèle de loi de probabilité (cf. chap. 6), un modèle de dépendances statistiques entre les variables analysées (cf. chap. 3), etc.

Exemple 8.1. Soit à étudier la loi de probabilité d'une variable aléatoire continue ξ à une dimension dont la densité de probabilité $f_\xi(x)$ est inconnue et supposons que l'analyse préliminaire des données initiales (8.1) (effectuée à l'aide des méthodes des chapitres 10 et 11) nous amène à conclure que cette loi peut être décrite par le *modèle normal* (cf. n° 6.1.5). Dans ce cas,

$$\Theta = (\theta^{(1)}, \theta^{(2)}), \text{ où } \theta^{(1)} = a = \mathbf{E}\xi \text{ et } \theta^{(2)} = \sigma^2 = \mathbf{V}\xi;$$

$$\mathbf{M}(x; \Theta) = f_\xi(x) - \frac{1}{\sqrt{2\pi\theta^{(2)}}} e^{-\frac{(x-\theta^{(1)})^2}{2\theta^{(2)}}}$$

et l'on peut prouver (cf., par exemple, [114]) que toute l'information sur les paramètres $\theta^{(1)}$ et $\theta^{(2)}$ (et, donc, sur le modèle) est concentrée

dans deux statistiques : $\bar{x}(n)$ et $s^2(n)$, où

$$\bar{x}(n) = \frac{1}{n} \sum_{i=1}^{n} x_i ; \tag{8.3}$$

$$s^2(n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}(n))^2. \tag{8.4}$$

On a montré plus haut (cf. § 7.2) qu'on avait toutes les raisons d'utiliser les statistiques (8.3) et (8.4) comme approximations (estimations) des paramètres a et σ^2, puisque ces estimations convergent stochastiquement lorsque $n \to \infty$ vers les valeurs véritables de a et σ^2. Mais le problème de savoir si ces estimations sont *les meilleures* reste pour l'instant ouvert. Avant de le discuter, nous aurons besoin d'introduire quelques notions.

8.1.2. Statistiques, estimations statistiques et leurs principales propriétés. On appelle *statistique* toute fonction $\gamma(X_1, \ldots, X_n)$ des observations $X_1, \ldots, X_n$ de la variable aléatoire multidimensionnelle ξ. Nous avons déjà eu affaire au chapitre 5 aux statistiques : la moyenne empirique (5.20′) ; la matrice des covariances empirique (5.36′) ; les coefficients empiriques de dissymétrie $\hat{\beta}_1(n)$ et d'aplatissement $\hat{\beta}_2(n)$ (formules (5.33′) et (5.34′)), la fonction de répartition empirique $\hat{F}^{(n)}(X)$ (formule (5.12) et la densité de probabilité $\hat{f}^{(n)}(X)$ (formule (5.23)).

La statistique $\hat{\Theta}$ qui sert à approcher le paramètre inconnu Θ s'appelle *estimation statistique*. Ainsi, par exemple, les statistiques $\bar{X}(n)$, $S(n)$, $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$ peuvent être traitées comme les estimations statistiques respectivement des paramètres $M = \mathbf{E}\xi$, $\Sigma = \mathbf{E}[(\xi - M)(\xi - M)']$, β_1 et β_2, puisque, en vertu du § 7.2, toutes ces statistiques convergent stochastiquement vers les valeurs exactes des paramètres respectifs lorsque $n \to \infty$.

Attirons l'attention du lecteur sur le fait qu'en parlant des statistiques et des estimations statistiques nous nous plaçons dans l'interprétation *hypothétique* de l'échantillon (8.1) (cf. note de la page 114), c'est-à-dire que $X_1, \ldots, X_n$ sont les n valeurs de ξ que *nous aurions pu obtenir* en réalisant n fois l'expérience aléatoire (ou n observations indépendantes) dans le système de conditions donné. Donc, les *statistiques et les estimations statistiques sont toutes des variables aléatoires* : au passage d'un échantillon à un autre (même parent) les *valeurs concrètes* de l'estimation statistique, calculées à l'aide de la formule (8.3) (c'est-à-dire les valeurs obtenues en portant dans (8.3) des valeurs concrètes de l'argument), feront l'objet

d'un éparpillement imprévisible. En fait, les valeurs de l'estimation statistique calculées sur des échantillons différents seront dispersées aléatoirement mais devront (si notre estimation est « bonne ») se rassembler autour de la véritable valeur du paramètre estimé.

Le problème est donc de formuler les critères que doivent remplir ces estimations pour être sûres dans un certain sens. Plus exactement ces estimations doivent être *convergentes*, *sans biais* et *efficaces*.

8.1.3. Estimation convergente. On dit qu'une estimation $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ d'un paramètre inconnu θ est *convergente* si elle tend stochastiquement vers la valeur à estimer θ lorsque $n \to \infty$, c'est-à-dire si $P\{|\hat{\theta} - \theta| > \xi\} \to 0$ lorsque $n \to \infty$ pour tout $\xi > 0$ aussi petit que l'on veut (si le paramètre Θ est vectoriel, la convergence de l'estimation vectorielle $\hat{\Theta}$ implique celle de toutes les composantes de $\hat{\Theta}$).

Toutes les estimations mentionnées plus haut ($\bar{X}(n)$, $S^2(n)$, $\hat{F}^{(n)}(X)$, $\hat{f}^{(n)}(X)$, etc.) sont, comme on l'a montré dans le § 7.2, des estimations convergentes des paramètres respectifs.

D'une part, la convergence est indispensable pour que cette estimation ait un sens pratique (puisque dans le cas contraire l'accroissement du volume de l'information initiale ne nous « rapprochera pas de la vérité »). Cette propriété doit donc être vérifiée d'emblée.

De l'autre, la convergence est une propriété asymptotique (par rapport au nombre d'observations n), c'est-à-dire qu'elle ne peut jouer que pour des échantillons de grande taille qui ne sont pas envisagés en pratique. Par ailleurs, dans la plupart des cas, on peut proposer plusieurs estimations convergentes d'un même paramètre. Par exemple, les estimations $\hat{\theta}_1 = \bar{x}(n)$ et $\hat{\theta}_2 = (x_{\min}(n) + x_{\max}(n))/2$ sont des estimations convergentes de la moyenne $\theta = E\xi$ (si elle existe) d'une variable aléatoire symétrique ξ ($\bar{x}(n)$ est la moyenne empirique calculée à l'aide de la formule (8.3), $x_{\min}(n)$ et $x_{\max}(n)$ respectivement les valeurs minimale et maximale des observations de ξ).

Tout ceci montre que la convergence ne suffit pas à elle seule à caractériser entièrement la fiabilité de l'estimation. Il faut envisager deux autres propriétés.

8.1.4. Estimation sans biais. On dit qu'une estimation $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ est *sans biais* si pour tout n sa moyenne $E\hat{\theta}$ est égale à la valeur théorique du paramètre inconnu θ, c'est-à-dire $E\hat{\theta} = \theta$ (si le paramètre Θ est vectoriel, la propriété d'une estimation vectorielle $\hat{\Theta}$ d'être sans biais implique celle de toutes les composantes de $\hat{\Theta}$).

Voyons, par exemple, si les estimations (8.3) et (8.4) sont des estimations sans biais des paramètres $a = \mathbf{E}\xi$ et $\sigma^2 = \mathbf{V}\xi$ de la loi normale (cf. exemple 8.1):

$$\mathbf{E}\bar{x}(n) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\,x_i = \frac{1}{n}\sum_{i=1}^{n} a = a;$$

$$\mathbf{E}s^2(n) = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}((x_i-a)-(\bar{x}-a))^2\right] =$$

$$= \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}(x_i-a)^2 - 2(\bar{x}-a)\sum_{i=1}^{n}(x_i-a) + \right.$$

$$\left. + \sum_{i=1}^{n}(\bar{x}-a)^2\right] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}(x_i-a)^2 - 2(\bar{x}-a)(n\bar{x}-na) + \right.$$

$$\left. + n(\bar{x}-a)^2\right] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}(x_i-a)^2 - n(\bar{x}-a)^2\right] =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{E}(x_i-a)^2 - n\cdot\mathbf{E}(\bar{x}-a)^2\right] =$$

$$= \frac{1}{n}\left(n\sigma^2 - n\cdot\frac{\sigma^2}{n}\right) = \sigma^2\left(1-\frac{1}{n}\right)$$

(dans le calcul de $\mathbf{E}s^2(n)$ on s'est servi du fait que si les observations $x_1, x_2, \ldots, x_n$ sont indépendantes, équiréparties et de variance σ^2, alors

$$\mathbf{V}\left(\frac{x_1+\ldots+x_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}x_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n};$$

cf. propriétés b) et d) de la variance au n° 5.6.3).

On constate que $\bar{x}(n)$ est une estimation sans biais du paramètre a, alors que l'estimation $s^2(n)$ du paramètre σ^2 admet un biais négatif égal à σ^2/n.

Contrairement à la convergence l'absence de biais caractérise les propriétés « *préasymptotiques* » de l'estimation, c'est-à-dire est une caractéristique de ses bonnes propriétés pour toute taille *finie* de l'échantillon. L'absence de biais élimine l'erreur systématique d'estimation qui, d'une façon générale, dépend de la taille n de l'échantillon et, en cas de convergence de l'estimation, tend en principe vers 0 lorsque $n \to \infty$. Si l'on arrive à déterminer la valeur du biais, on peut facilement l'éliminer. Ainsi, dans notre exemple,

pour éliminer le biais, il suffit de passer à l'estimation

$$\bar{s}^2(n) = \frac{n}{n-1} s^2(n)$$

qui de toute évidence est sans biais. De ce qui précède, il suit que l'absence de biais (sous réserve que la convergence ait lieu) est particulièrement importante lorsque le nombre d'observations est petit.

8.1.5. Estimation efficace. Soient données deux estimations vectorielles convergentes et sans biais $\hat{\Theta}_1(X_1, \ldots, X_n)$ et $\hat{\Theta}_2(X_1, \ldots, X_n)$ d'un paramètre vectoriel inconnu Θ. Pour interpréter

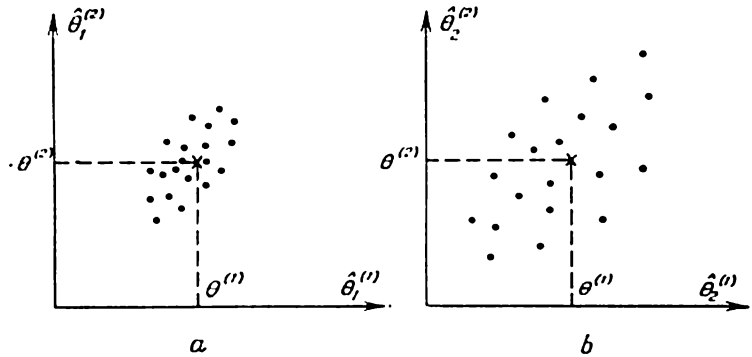


Fig. 8.1. Deux procédés d'estimation sans biais convergente d'un paramètre vectoriel $\Theta = (\theta^{(1)}, \theta^{(2)})$ d'efficacité différente : sur *a* l'estimation est plus efficace que sur *b*

géométriquement cet exemple, on admettra que le paramètre vectoriel est de dimension 2. Pour analyser les propriétés de deux estimations en concurrence, on réalisera ici 20 estimations du paramètre inconnu $\Theta = (\theta^{(1)}, \theta^{(2)})'$ à l'aide des deux méthodes préconisées. A cet effet, calculons les valeurs $\hat{\Theta}_{1i}$ et $\hat{\Theta}_{2i}$ ($i = 1, 2, \ldots, 20$) en portant dans les fonctions $\hat{\Theta}_1$ et $\hat{\Theta}_2$ le i-ième échantillon de taille n, autrement dit, prélevons le premier échantillon $X_{11}, X_{12}, \ldots, X_{1n}$ et portons ces observations dans les fonctions $\hat{\Theta}_1$ et $\hat{\Theta}_2$ pour obtenir le premier couple d'estimations $\hat{\Theta}_{11}$ et $\hat{\Theta}_{21}$; tirons, ensuite, un deuxième échantillon $X_{21}, \ldots, X_{2n}$ et portons ces observations dans les *mêmes* fonctions $\hat{\Theta}_1$ et $\hat{\Theta}_2$ pour obtenir le deuxième couple d'estimations $\hat{\Theta}_{12}$ et $\hat{\Theta}_{22}$, et ainsi de suite. Sur l'axe horizontal de la figure 8.1 sont portées la première composante $\theta^{(1)}$ du paramètre inconnu et les premières composantes de ses deux estimations

($\hat{\theta}_{1i}^{(1)}$ sur la figure 8.1, *a* et $\hat{\theta}_{2i}^{(1)}$ sur la figure 8.1, *b*) et sur l'axe vertical, la deuxième composante $\hat{\theta}^{(2)}$ du paramètre inconnu et les deuxièmes composantes de ses deux estimations ($\hat{\theta}_{1i}^{(2)}$ sur la figure 8.1, *a* et $\hat{\theta}_{2i}^{(2)}$ sur la figure 8.1, *b*). Donc, la position relative du point ($\hat{\theta}_{1i}^{(1)}$, $\hat{\theta}_{1i}^{(2)}$) et de la croix ($\theta^{(1)}$, $\theta^{(2)}$) sur la figure 8.1, *a* donne une idée de la proximité de l'estimation $\hat{\Theta}_{1i}$ obtenue par la première méthode à l'aide du *i*-ième échantillon, par rapport à la valeur théorique, du paramètre inconnu Θ (la figure 8.1, *b* représente la même situation pour la deuxième méthode d'estimation). La plus forte concentration des estimations acquises par la première méthode autour de la valeur théorique témoigne visiblement de la plus grande efficacité de l'estimation $\hat{\Theta}_1$ par rapport à $\hat{\Theta}_2$. C'est précisément cette mesure de la dispersion des valeurs estimées $\hat{\Theta}$ autour de la valeur théorique de Θ dans l'espace *k*-dimensionnel respectif qui est posée à la base de la définition de l'efficacité d'une estimation. On dit qu'une *estimation* $\hat{\Theta}$ *d'un paramètre* Θ *est efficace si de toutes les estimations de* Θ *elle est celle dont la mesure de la dispersion aléatoire par rapport à la valeur théorique de* Θ *est la plus petite.* L'efficacité est une propriété décisive qui définit la qualité d'une estimation et qui, généralement, n'implique pas l'absence de biais.

Reste à préciser comment varie le degré de dispersion aléatoire des valeurs de l'estimation $\hat{\Theta}$ autour de la valeur théorique du paramètre Θ.

Lorsque $\hat{\Theta}$ est un *scalaire* (c'est-à-dire que la dimension *k* de l'estimation est égale à 1), pour mesurer cette dispersion, on prend la moyenne du carré de l'écart, c'est-à-dire la quantité $\sigma^2(\hat{\Theta}) = \mathbf{E}(\hat{\Theta} - \Theta)^2$ qui, pour les estimations sans biais, coïncide avec leur dispersion, puisque dans ce cas $\mathbf{V}\hat{\Theta} = \mathbf{E}(\Theta - \mathbf{E}\hat{\Theta})^2 = \mathbf{E}(\hat{\Theta} - \Theta)^2$.

Lorsque $\hat{\Theta}$ est un *vecteur* (c'est-à-dire que $k \geqslant 2$), pour mesurer l'écart par rapport à la valeur théorique du paramètre vectoriel Θ, on prend généralement la matrice des covariances de l'estimation $\hat{\Theta}$, c'est-à-dire une matrice symétrique définie positive de dimension $k \times k$ que l'on désigne par $\Sigma(\hat{\Theta})$. Une *estimation* $\hat{\Theta}_1$ *du paramètre* Θ *est dite moins efficace qu'une estimation* $\hat{\Theta}_2$ *si existent les matrices des covariances* $\Sigma(\hat{\Theta}_1)$ *et* $\Sigma(\hat{\Theta}_2)$ *et si la matrice* $\Delta\Sigma = \Sigma(\hat{\Theta}_1) - \Sigma(\hat{\Theta}_2)$ *est semi-définie positive.*

Si les estimations sont vectorielles, il existe des cas où, malgré

l'existence des matrices $\Sigma(\hat{\Theta}_1)$ et $\Sigma(\hat{\Theta}_2)$, il est impossible de dire laquelle est plus efficace au sens indiqué ci-dessus. Cependant, on peut lever cette indétermination si pour mesurer l'écart d'une estimation vectorielle sans biais $\hat{\Theta}$ par rapport à la valeur théorique du paramètre Θ on considère non pas la matrice des covariances $\Sigma(\hat{\Theta})$ mais son déterminant $\det \Sigma(\hat{\Theta})$ (la *dispersion généralisée*, cf. n° 5.6.7) ou sa trace $\mathrm{Tr}\, \Sigma(\hat{\Theta})$.

8.2. Fonction de vraisemblance. Quantité d'information contenue dans n observations indépendantes par rapport à la valeur inconnue du paramètre

Supposons que (8.1) est un échantillon constitué de n observations indépendantes p-dimensionnelles. La loi de probabilité de la variable aléatoire p-dimensionnelle ξ est décrite par une fonction $f(X, \Theta)$ dépendant d'un paramètre inconnu Θ : par $f(X, \Theta)$ on comprend la probabilité $P\{\xi = X\}$ si ξ est discrète et la valeur de la densité de probabilité en X si ξ est continue. Si l'on envisage l'échantillon (8.1) au sens hypothétique, chaque échantillon *concret* $(X_1^*, X_2^*, \ldots, X_n^*)$ est représenté par un point de l'espace à $(p \times n)$ dimensions des échantillons $X_1, X_2, \ldots, X_n$ et il y a lieu de parler de la loi conjointe du vecteur $\mathbf{X} = (X_1, \ldots, X_n)$. Vu que dans l'optique hypothétique, $X_1, X_2, \ldots, X_n$ sont des variables aléatoires indépendantes équiréparties, pour tout ensemble *donné* de valeurs $X_1^*, \ldots, X_n^*$, leur densité conjointe sera

$$L(X_1^*, X_2^*, \ldots, X_n^*; \Theta) = f(X_1^*; \Theta) \cdot f(X_2^*; \Theta) \cdot \ldots \cdot f(X_n^*; \Theta). \tag{8.5}$$

Donc, la fonction $L(\mathbf{X}^*, \Theta)$ définie par (8.5) nous donne la probabilité d'obtenir par un tirage d'un échantillon de taille n les observations $X_1^*, \ldots, X_n^*$ (ou une quantité proportionnelle à la probabilité d'obtenir des valeurs empiriques situées au voisinage immédiat du point $\mathbf{X}^*$ dans le cas continu). Donc, plus la valeur de $L(\mathbf{X}^*, \Theta)$ est élevée et plus est *vraisemblable* le système d'observations $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ pour une valeur donnée du paramètre Θ. D'où le nom de *fonction de vraisemblance* donné à la fonction L.

Selon la position du problème et les objectifs de la recherche, la fonction de vraisemblance peut être traitée soit comme une fonction du paramètre Θ (pour des observations fixes données $X_1^*, \ldots, X_n^*$), soit comme une fonction des valeurs courantes des observations $X_1, \ldots, X_n$ (pour une valeur fixe donnée du paramètre Θ), soit enfin comme une fonction de $\mathbf{X}$ et Θ.

Il serait intéressant d'étudier les variations de la probabilité en fonction de celles du paramètre Θ. Il est évident que plus cette dépendance est forte et plus importante est l'information comprise dans X sur Θ. Signalons que par information comprise dans la variable aléatoire X sur le paramètre inconnu Θ, on comprend le degré de réduction de l'indétermination relative à la valeur inconnue de Θ après les observations effectuées sur la variable X. Si la valeur observée X^* de la variable aléatoire X permet de remonter presque sûrement (c'est-à-dire avec la probabilité 1) à la valeur exacte du paramètre Θ, c'est que X (ou sa valeur observée X^*) renferme le maximum d'information sur le paramètre Θ. Réciproquement, si la répartition (8.5) de la variable aléatoire X est la même pour toutes les valeurs du paramètre Θ, nous n'avons alors aucune raison de tirer une quelconque conclusion sur Θ d'après les résultats des observations de cette variable aléatoire (ceci exprime que l'observation contient une information nulle sur la valeur du paramètre inconnu). La sensibilité de la variable aléatoire X au paramètre Θ peut être mesurée par la grandeur de la variation de la répartition de X lorsque Θ varie. La caractéristique la plus souvent utilisée et qui sert à mesurer la distance entre les répartitions (8.5) pour deux valeurs différentes de Θ est la *quantité d'information de Fisher* (contenue dans les observations $X = (X_1, X_2, \ldots, X_n)$) définie comme suit

$$\mathbf{I}(\theta;\ X) = \mathbf{E}\left[\left(\frac{d \ln L}{d\theta}\right)^2\right] = \int \left(\frac{d}{d\theta} \ln L(X,\ \theta)\right)^2 L(X;\ \theta)\ dX \quad (8.6)$$

pour un paramètre scalaire θ (c'est-à-dire pour un paramètre Θ de dimension 1).

Vu que les observations $X_1, X_2, \ldots, X_n$ sont indépendantes et équiréparties, il vient

$$\mathbf{I}(\theta;\ X) = n \int \left(\frac{d \ln f(X;\theta)}{d\theta}\right)^2 f(X;\ \theta)\, dx = n\mathbf{I}(\theta;\ X). \quad (8.6')$$

Si le paramètre $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ est de dimension $k \geqslant 2$, au lieu de la quantité d'information (8.6), on considère la *matrice d'information de Fisher* $\mathbf{I}(\Theta, X)$ de dimension $k \times k$, d'éléments

$$I_{ij}(\Theta, X) = \mathbf{E}\left[\left(\frac{\partial \ln L}{\partial \theta^{(i)}} \cdot \frac{\partial \ln L}{\partial \theta^{(j)}}\right)\right]. \quad (8.7)$$

Ces notions ont été introduites par Fisher dans les années 20.

La formule (8.6) nous permet de calculer sans peine la quantité d'information $I(\theta;\ X)$ contenue dans une observation du paramètre θ dans bien des cas.

1. La variable aléatoire à une dimension $X = x$ suit une loi normale de paramètres (a, σ^2) et de densité $\varphi(x;\ a, \sigma^2)$ (cf. n° 6.1.5), où la moyenne $a = \theta$ est un paramètre inconnu et la variance σ^2 est

connue. Alors

$$I(a;\ x)=\int_{-\infty}^{\infty}\left(\frac{d\ln\varphi(x;\ a,\ \sigma^2)}{da}\right)^2\varphi(x;\ a,\ \sigma^2)\,dx=$$

$$=\int_{-\infty}^{\infty}\left(\frac{x-a}{\sigma^2}\right)^2\varphi(x;\ a,\ \sigma^2)\,dx=\frac{1}{\sigma^2}. \qquad (8.8)$$

Ce résultat admet l'interprétation naturelle suivante: plus la variance σ^2 est élevée, plus la dispersion des valeurs observées de X est grande et moins une observation contient d'information sur la moyenne de X.

2. La variable aléatoire à une dimension $X = x$ suit une loi normale de paramètres $(a,\ \sigma^2)$, de densité $\varphi(x;\ a,\ \sigma^2)$ (cf. n° 6.1.5), où la moyenne a est connue et la variance $\sigma^2 = \theta$ est un paramètre inconnu. Alors

$$I(\sigma^2;\ x)=\int_{-\infty}^{\infty}\left(\frac{d\ln\varphi(x;\ a,\ \sigma^2)}{d\sigma^2}\right)^2\varphi(x;\ a,\ \sigma^2)\,dx=$$

$$=\int_{-\infty}^{\infty}\left[\frac{(x-a)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right]^2\varphi(x;\ a,\ \sigma^2)\,dx=\frac{1}{2\sigma^4} \qquad (8.9)$$

3. La variable aléatoire à une dimension $X = x$ suit une loi gamma de paramètres $(a,\ b)$ dont a est connu et b inconnu (cf. n° 6.2.5). Alors

$$I(b;\ x)=\int_{0}^{\infty}\left(\frac{d\ln f_{\gamma(a,\ b)}(x)}{db}\right)\cdot f_{\gamma(a,\ b)}(x)\,dx=\frac{a}{b^2}. \qquad (8.10)$$

8.3. Inégalité de Rao-Cramer-Fréchet et mesure de l'efficacité des estimations

Au n° 8.1.5 nous avons défini l'efficacité d'une estimation $\hat{\Theta}$ d'un paramètre inconnu Θ comme la moyenne du carré de l'écart de $\hat{\Theta}$ par rapport à la valeur théorique de Θ, soit $\mathbf{E}(\hat{\Theta}-\Theta)^2$. Cette définition appelle la question suivante: n'est-il pas possible de décrire les bornes de cette efficacité, c'est-à-dire le minimum (sur toutes les estimations $\hat{\Theta}$) de la moyenne $\mathbf{E}(\hat{\Theta}-\Theta)^2$ qu'il est impossible d'améliorer? Ce minimum nous servirait de point de départ pour établir une échelle absolue pour mesurer l'efficacité des estimations. La réponse à cette question nous est fournie par l'inégalité de Rao-Cramer-Fréchet connue encore sous le nom d'*inégalité d'information*.

Considérons la classe de toutes les estimations $\hat{\theta}$ d'un paramètre *scalaire* θ dont dépend la densité de probabilité $f(X;\theta)$ de la population générale étudiée. Posons

$$\mathrm{E}\hat{\theta} = \int \hat{\theta}(X_1, \ldots, X_n)\, L(X_1, \ldots, X_n; \theta)\, dX_1 \ldots \ldots dX_n = \theta + b_{\hat{\theta}}(\theta), \tag{8.11}$$

c'est-à-dire que la quantité $b_{\hat{\theta}}(\theta)$ nous donne le biais de l'estimation $\hat{\theta}$ (il est évident que $b_{\hat{\theta}}(\theta) \equiv 0$ si $\hat{\theta}$ est sans biais).

Si la densité de probabilité $f(X;\theta)$ satisfait *certaines conditions de régularité* (par rapport au paramètre θ), plus exactement, si

a) le domaine des valeurs possibles de la variable aléatoire dans lequel $f(X;\theta) \neq 0$ *ne dépend pas de* θ;

b) la dérivation par rapport à θ est licite sous le signe d'intégration dans la formule (8.11) et dans l'identité $\int L(X_1, \ldots, X_n; \theta)\, dX_1 \ldots dX_n \equiv 1$;

c) la quantité $I(\theta; X)$ définie par (8.6') n'est pas nulle, alors pour toute estimation $\hat{\theta}$ du paramètre θ, on a l'inégalité

$$\mathrm{E}(\hat{\theta} - \theta)^2 \geqslant \frac{\left(1 + \frac{db_{\hat{\theta}}(\theta)}{d\theta}\right)^2}{n \cdot \mathrm{E}\left[\left(\frac{d \ln f(X;\theta)}{d\theta}\right)^2\right]}, \tag{8.12}$$

ou, ce qui revient au même,

$$\mathrm{E}(\hat{\theta} - \theta)^2 \geqslant \frac{\left(1 + \frac{db_{\hat{\theta}}(\theta)}{d\theta}\right)^2}{nI(\theta; X)}. \tag{8.12'}$$

L'inégalité (8.12) se généralise à un paramètre $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ de dimension $k \geqslant 2$. Dans ce cas, sous les mêmes conditions de régularité a), b) et c), la matrice

$$\Sigma(\hat{\Theta}) - \frac{1}{n}\mathbf{I}^{-1}(\Theta, X) \tag{8.13}$$

est semi-définie positive pour toute estimation vectorielle *sans biais* $\hat{\Theta}$ du paramètre Θ: $\Sigma(\hat{\Theta})$ est la matrice des covariances de l'estimation vectorielle $\hat{\Theta} = (\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(k)})$ et $I^{-1}(\Theta, X)$ la matrice inverse de la matrice de Fisher définie par les relations (8.7) pour $\mathbf{X} = X$ (c'est-à-dire pour l'unique observation X).

Les inégalités d'information (8.12), (8.12') et (8.13) nous permettent d'introduire une mesure de l'efficacité des estimations sur la classe des populations générales régulières (au sens des conditions

a), b) et c)). Il semble, en particulier, naturel de mesurer l'efficacité d'une estimation scalaire sans biais $\hat{\theta}$ de θ par le rapport $e(\hat{\theta})$ de la plus petite valeur de la variance de $\hat{\theta}$, qui est définie par le second membre de (8.12), à la variance de $\hat{\theta}$, soit

$$e(\hat{\theta}) = \frac{1}{n \cdot I(\theta; X)} : \mathrm{E}(\hat{\theta} - \theta)^2. \tag{8.14}$$

Calculons l'efficacité de certaines estimations des paramètres a et σ^2 dans les conditions de l'exemple 8.1.

1. Prenons pour estimation de la moyenne a d'une variable aléatoire normale la moyenne arithmétique empirique, c'est-à-dire posons

$$\hat{a} = \bar{x}(n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Les observations x_i étant indépendantes et équiréparties, on a

$$\mathrm{E}\hat{a} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}x_i = \frac{1}{n} \sum_{i=1}^{n} a = a;$$

$$\mathrm{V}\hat{a} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{V}x_i = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Comme $I(a, x) = 1/\sigma^2$ (cf. (8.8)), la formule (8.14) nous donne $e(\bar{x}(n)) = 1$, c'est-à-dire que l'estimation $\bar{x}(n)$ d'une moyenne normale est *inaméliorable*.

2. Prenons pour estimation de la variance σ^2 d'une variable aléatoire normale « corrigée » la variance empirique

$$\bar{s}^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}(n))^2.$$

On a vu plus haut (cf. n° 8.1.4) que $\mathrm{E}\bar{s}^2(n) = \sigma^2$, c'est-à-dire que $\bar{s}^2(n)$ est une estimation sans biais de σ^2. Les calculs nous donnent (cf. par exemple [40])

$$\mathrm{V}\bar{s}^2(n) = \frac{n^2}{(n-1)^2} \cdot \mathrm{V}s^2(n) = \frac{n^2}{(n-1)^2} \cdot \frac{2(n-1)}{n^2} \sigma^4 = \frac{2}{n-1} \sigma^4.$$

Comme $I(\sigma^2, x) = \frac{1}{2\sigma^4}$ (cf. (8.9)), la formule (8.14) nous donne $e(\bar{s}^2(n)) = (n-1)/n$, c'est-à-dire que l'estimation $\bar{s}^2(n)$ n'est pas efficace bien qu'elle soit proche de σ^2 pour les échantillons de grande taille. On démontre par ailleurs que si pour estimation de σ^2 on re-

tient la statistique

$$s_0^2(n) = \frac{1}{n} \sum_{i=1}^{n} (x_j - a)^2,$$

ce qui est licite, puisque la moyenne a est *connue*, alors elle est efficace.

Signalons en conclusion que l'*inégalité d'information n'est valable que sur la classe des populations générales régulières* (*au sens des conditions* a), b) et c) *du* § 8.3). Si, en particulier, le domaine des valeurs possibles de la variable aléatoire, pour lesquelles la densité $f(x; \theta)$ est strictement positive, dépend du paramètre θ, alors l'inégalité d'information est mise en défaut. Présentent de telles densités non régulières la répartition uniforme (dont les paramètres sont les bornes de l'intervalle de variation de la variable aléatoire, cf. n° 6.1.7), la répartition exponentielle de paramètre de localisation θ, c'est-à-dire la répartition de densité

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & \text{pour } x \geqslant \theta; \\ 0 & \text{pour } x < \theta. \end{cases}$$

Si en passant outre le fait que cette densité ne vérifie pas les conditions a), b) et c), on calcule à l'aide de la formule (8.6) la quantité d'information contenue dans n observations indépendantes, on trouve que $I(\theta; x_1, \ldots, x_n) = n$. Donc, en vertu de l'inégalité d'information (8.12), on aurait dû conclure que la variance d'aucune estimation $\hat{\theta}$ de θ ne peut être strictement inférieure à $\frac{1}{n}$. En même temps, il est aisé de calculer (cf. n° 8.6.5) la moyenne et la variance

$$\mathbf{E}\hat{\theta} = \theta \quad \text{et} \quad \mathbf{V}\hat{\theta} = \frac{1}{n^2}$$

de l'estimation

$$\hat{\theta} = x_{\min}(n) - \frac{1}{n}, \tag{8.15}$$

où $x_{\min}(n)$ désigne comme toujours la valeur minimale de l'échantillon $x_1, \ldots, x_n$.

Donc, si l'on se sert de la formule (8.14) pour mesurer l'efficacité de l'estimation (8.15), on constate que l'efficacité de l'estimation $\hat{\theta}$ n'est pas simplement supérieure à l'unité, mais elle tend vers l'infini lorsque la taille n de l'échantillon croît (puisque $e(\hat{\theta}) = \frac{1}{n} : \frac{1}{n^2} = n$). De telles estimations sont dites « *superefficaces* ».

Remarque sur les variables aléatoires discrètes. Tous les résultats exposés ci-dessus (notion de quantité d'information, inégalité d'in-

formation, mesure de l'efficacité d'une estimation) se généralisent aux variables discrètes avec les mêmes contraintes a), b) et c) moyennant quelques modifications évidentes : la densité $f(x;\theta)$ doit être remplacée par les probabilités $p_i(\theta) = P\{\xi = x_i^0 \mid \theta\}$ et l'intégration, par une sommation sur toutes les valeurs possibles de la variable aléatoire discrète. Donc, les analogues discrets de la quantité d'information (8.6′) et de l'inégalité d'information (8.12) seront

$$I(\theta;\, x_1, \ldots, x_n) = nI(\theta;\, x) = n\sum_i \left(\frac{d\ln p_i(\theta)}{d\theta}\right)^2 p_i(\theta); \tag{8.16}$$

$$V\hat{\theta} \geqslant \frac{\left(1+\frac{d\,b(\theta)}{d\theta}\right)^2}{n\sum\limits_i \left(\frac{d\ln p_i(\theta)}{d\theta}\right)^2 p_i(\theta)}. \tag{8.17}$$

Considérons à titre d'exemple une variable aléatoire de Poisson (cf. n° 6.1.2), c'est-à-dire que

$$p_i(\lambda) = P\{\xi = i \mid \lambda\} = \frac{\lambda^i}{i!}e^{-\lambda}, \quad i = 0, 1, 2, \ldots$$

$$I(\lambda;\, x) = \sum_{i=0}^{\infty}\left(\frac{d\ln p_i(\lambda)}{d\lambda}\right)^2 p_i = \sum_{i=0}^{\infty}\left(\frac{i}{\lambda}-1\right)^2 p_i = \frac{1}{\lambda},$$

autrement dit, la variance de toute estimation sans biais $\hat{\lambda}$ de λ ne peut être strictement inférieure à $\frac{1}{n}\mathrm{I}(\lambda, x) = \frac{\lambda}{n}$. Si, pour estimation de λ, on prend la moyenne empirique $\bar{x}(n)$, alors

$$\mathrm{E}\hat{\lambda} = \mathrm{E}\left(\frac{x_1+\ldots+x_n}{n}\right) = \frac{n\mathrm{E}x}{n} = \lambda;$$

$$\mathrm{V}\hat{\lambda} = \mathrm{V}\left(\frac{x_1+\ldots+x_n}{n}\right) = \frac{1}{n^2}n\mathrm{V}x = \frac{\lambda}{n}.$$

Donc, l'estimation $\lambda = \bar{x}(n)$ du paramètre λ dans la répartition de Poisson est efficace.

8.4. Propriétés asymptotiques des estimations

Toute estimation $\hat{\theta}(X_1, \ldots, X_n)$ traitée comme une fonction des résultats « hypothétiques » des observations est une *variable aléatoire* dont les propriétés sont, par conséquent, définies par sa fonction de répartition $F_{\hat{\theta}}(u;\, n)$. Vu que l'estimation est construite à l'aide d'un échantillon de taille n finie, sa fonction de répartition dépendra généralement de n, ce qui d'ailleurs est exprimé dans sa notation. Mais si la définition pratique des lois de probabilité des

estimations pour des échantillons de taille n finie est très compliquée dans la plupart des cas, il est, par contre, bien plus simple de calculer leur loi *asymptotique* (moyennant une normalisation adéquate). Soit, en particulier, $\hat{\theta}$ une estimation de θ et supposons qu'il existe une fonction de répartition $F_{\hat{\theta}_{\text{norm}}}(u)$ dérivable et continue, telle qu'en tout point u l'on ait

$$P\{\sqrt{n}(\hat{\theta}-\theta) < u\} \to F_{\hat{\theta}_{\text{norm}}}(u) \text{ lorsque } n \to \infty. \tag{8.18}$$

Dans ce cas, la fonction $F_{\hat{\theta}}(u) = \tilde{F}_{\hat{\theta}_{\text{norm}}}(\sqrt{n}\,u)$ s'appelle *fonction de répartition asymptotique* de l'estimation $\hat{\theta}$ et les valeurs

$$b_{\text{as}}(\theta) = \frac{1}{\sqrt{n}} \int u \tilde{f}_{\hat{\theta}_{\text{norm}}}(u)\, du$$

et

$$\sigma^2_{\text{as}}(\hat{\theta}) = \frac{1}{\sqrt{n}} \int \left[u - \int v \cdot \tilde{f}_{\hat{\theta}_{\text{norm}}}(v)\, dv\right]^2 \tilde{f}_{\hat{\theta}_{\text{norm}}}(u)\, du$$

respectivement *biais asymptotique* et *variance asymptotique* de cette estimation.

Si le biais asymptotique est nul, l'estimation est dite *asymptotiquement sans biais*. Une estimation asymptotiquement sans biais n'est pas sans biais au sens ordinaire et, réciproquement, une estimation sans biais n'est pas nécessairement asymptotiquement sans biais. Mais en pratique la propriété d'être sans biais est plus forte et les estimations sans biais sont généralement asymptotiquement sans biais. De même, il n'est pas obligatoire que σ^2_{as} tende vers la variance de l'estimation pour une normalisation convenable lorsque $n \to \infty$. La variance de l'estimation peut en effet ne pas exister. Pourtant, on ne peut affirmer que la variance asymptotique de l'estimation est toujours inférieure à la variance, bien que ceci ait généralement lieu en pratique.

Exemple 8.2 (la fréquence comme mesure de la probabilité). En vertu du théorème limite central (cf. § 7.3), la fréquence $\hat{p} = \frac{\nu}{n}$ d'un événement dans une série de n expériences suit une loi asymptotiquement normale de moyenne égale à la probabilité respective $p = \mathbf{E}\hat{p}$ et de variance $\sigma^2 = p(1-p)/n$, c'est-à-dire que

$$P\{(\hat{p} - p)/\sigma < u\} \to \Phi(u) \text{ lorsque } n \to \infty.$$

Si l'on traite $\hat{p}$ comme une estimation du paramètre p, on remarque que c'est une estimation sans biais et asymptotiquement sans biais dont la variance et la variance asymptotique sont égales à σ^2.

Si $\hat{\Theta}$ est une estimation vectorielle du vecteur Θ, on peut, par analogie avec (8.18), définir une fonction de répartition $F_{\hat{\theta}_{\text{norm}}}(U)$ de dimension k. Pour cela, il suffit dans (8.18) de remplacer $\hat{\theta}$, θ et u respectivement par les vecteurs $\hat{\Theta}$, Θ et U et de comprendre cette inégalité comme une inégalité réalisée simultanément pour toutes les composantes. Supposons par ailleurs que $\hat{\Theta}$ est une estimation asymptotiquement sans biais de Θ, c'est-à-dire que $\int U \cdot \tilde{f}_{\hat{\theta}_{\text{norm}}}(U)\, dU = 0$ et qu'existe la matrice des covariances

$$\Sigma = \left\| \int u^{(i)} u^{(j)} \tilde{f}_{\hat{\theta}_{\text{norm}}}(U)\, dU \right\|, \quad i, j = 1, \ldots, k.$$

Alors la matrice $n^{-1}\Sigma$ sera appelée matrice des covariances asymptotique de l'estimation Θ.

Définissons la notion d'*efficacité asymptotique d'une estimation.* Soient $\hat{\Theta}_1$ et $\hat{\Theta}_2$ deux estimations asymptotiquement sans biais différentes de Θ. On dit que l'estimation $\hat{\Theta}_1$ est asymptotiquement plus efficace que l'estimation $\hat{\Theta}_2$ si la variance asymptotique de $\hat{\Theta}_1$ est strictement inférieure à la variance asymptotique de $\hat{\Theta}_2$. S'agissant des estimations asymptotiquement sans biais *vectorielles*, on dit qu'une estimation $\hat{\Theta}_1$ est asymptotiquement plus efficace qu'une estimation $\hat{\Theta}_2$ si existent les matrices des covariances asymptotiques Σ_1 et Σ_2 de ces estimations et si, de plus, la matrice $\Sigma_2 - \Sigma_1$ est semi-définie positive. Si les estimations sont vectorielles, il est possible qu'on ne puisse dire laquelle des deux est la plus efficace malgré l'existence des matrices Σ_1 et Σ_2.

8.5. Notion d'estimation par intervalles. Construction des régions de confiance

Quand on calcule une estimation $\hat{\Theta}(X_1, \ldots, X_n)$ d'un paramètre θ à l'aide des observations $X_1, \ldots, X_n$, on sait pertinement que $\hat{\theta}$ n'est qu'une valeur *approchée* du paramètre θ même dans le cas où cette estimation est convergente (c'est-à-dire tend vers θ lorsque n croît), est sans biais (c'est-à-dire est confondue avec θ en moyenne) et efficace (c'est-à-dire présente les plus petits écarts aléatoires par rapport à θ). Une question se pose : quel est le plus grand écart entre la valeur approchée (l'estimation) et la valeur *véritable* de θ? En particulier, peut-on indiquer une quantité Δ telle que $|\hat{\theta} - \theta| < \Delta$ avec une probabilité proche de l'unité donnée à l'avance? Ou encore, peut-on indiquer un intervalle de la forme

$[\hat{\theta} - \Delta, \hat{\theta} + \Delta[$ qui contienne la valeur véritable de θ avec une probabilité donnée à l'avance proche de l'unité? Cette probabilité s'appelle *niveau de confiance*, l'intervalle $]\hat{\theta} - \Delta, \hat{\theta} + \Delta[$, *intervalle de confiance*. L'intervalle de confiance est *aléatoire* (c'est pourquoi il est question de la *probabilité de recouvrir* un point inconnu mais non aléatoire θ !) aussi bien par sa position ($\hat{\theta}$ n'est-elle pas aléatoire?) que par sa longueur (Δ est en général une fonction des données empiriques $X_1, \ldots, X_n$). La longueur de l'intervalle de confiance dépend essentiellement de la taille n de l'échantillon (il diminue lorsque n croît) et du niveau de confiance (il augmente lorsque ce niveau se rapproche de l'unité).

Toutes les définitions et notions de ce paragraphe se généralisent sans peine à un paramètre *vectoriel* $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ en remplaçant l'intervalle de confiance par une *région de confiance* dans l'espace à k dimensions correspondant (voir la forme de ces régions sur la figure 8.1, *a* et *b*).

8.6. Méthodes d'estimations statistiques des paramètres inconnus

Au paragraphe précédent nous avons vu comment utiliser les fonctions des observations initiales $X_1, \ldots, X_n$ pour estimer les paramètres inconnus et avons étudié leurs propriétés. Mais on ne sait encore pas comment combiner les résultats de ces observations pour construire les statistiques à l'aide desquelles on estimera de la meilleure façon (dans un sens précis) tel ou tel paramètre. On ne sait encore pas, par exemple, comment on a établi que $\bar{x}(n)$ et $s^2(n)$ (cf. (8.3), (8.4)) se prêtaient le mieux à l'estimation de la moyenne $a = \mathbf{E}\xi$ et de la variance $\sigma^2 = \mathbf{V}\xi$ d'une population générale normale. On ne sait, enfin, encore pas comment construire les intervalles et les régions de confiance pour les valeurs inconnues des paramètres.

On se propose dans ce paragraphe de décrire des méthodes qui nous permettront de répondre à toutes ces questions.

8.6.1. Méthode du maximum de vraisemblance. D'après cette méthode, toute estimation $\hat{\Theta}_{mv}$ du paramètre inconnu Θ sur le vu des observations $X_1, \ldots, X_n$ d'une variable aléatoire ξ (suivant une loi de probabilité de densité $f_\xi(X; \Theta)$ ou de probabilité $P\{\xi = X\}$) se définit à partir de la condition

$$L(X_1, \ldots, X_n; \hat{\theta}_{mv}) = \max_{\hat{\theta}} L(X_1, \ldots, X_n; \hat{\theta}, \quad (8.19)$$

où L est la fonction de vraisemblance définie par la relation (8.5).

Donc, l'*estimation du maximum de vraisemblance* $\hat{\Theta}_{mv}$ de Θ sur le vu des observations *indépendantes* $X_1, \ldots, X_n$ peut être écrite

formellement :

$$\hat{\theta}_{mv} = \operatorname{Arg\,max}_{\hat{\theta}} \prod_{i=1}^{n} f(X_i; \hat{\theta}). \tag{8.19'}$$

Le caractère naturel de cette définition des estimations statistiques résulte de la signification de la fonction de vraisemblance. En effet, par définition (cf. § 8.2), la fonction $L(X_1, \ldots, X_n; \Theta)$ est pour toute valeur fixe de Θ la mesure de la vraisemblance des

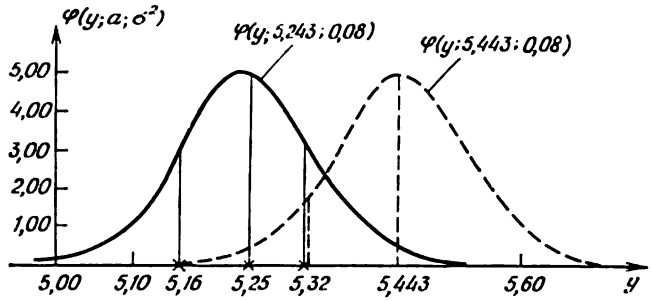


Fig. 8.2. Représentations graphiques de la densité normale pour deux valeurs distinctes du paramètre a

valeurs observées $X_1, \ldots, X_n$. C'est pourquoi, en se donnant des valeurs *concrètes* de $X_1, \ldots, X_n$, on peut voir pour quelles valeurs de Θ ces observations sont *plus vraisemblables* et choisir en fin de compte la valeur de $\hat{\Theta}_{mv}$, pour laquelle le système d'observations $X_1, \ldots, X_n$ est le *plus vraisemblable* (il est évident que $\hat{\Theta}_{mv}$ est une fonction de $X_1, \ldots, X_n$). Supposons, par exemple, qu'un salaire ξ suit une loi lognormale (cf. n° 6.1.6). Supposons que dans le but d'estimer la moyenne du logarithme du salaire $a = \mathbf{E}(\ln \xi)$ on ait fixé les salaires $x_1 = 190$ roubles, $x_2 = 175$ roubles et $x_3 = 205$ roubles de trois personnes prélevées au hasard dans la population générale étudiée. Portons les valeurs $y_i = \ln x_i$ $(i = 1, 2, 3)$ sur l'axe des valeurs possibles de la variable normale $\eta = \ln \xi$ et essayons de trouver une valeur $\hat{a}_{mv}$ du paramètre a de la loi normale $N(a, \sigma^2)$ telle que les observations y_1, y_2, y_3 soient *les plus vraisemblables*, plus exactement, telle que le produit des ordonnées de la densité $\varphi(y; a; \sigma^2)$ calculées aux points respectivement d'abscisses $y_1 = \ln 190 = 5{,}25$, $y_2 = \ln 175 = 5{,}16$ et $y_3 = \ln 205 = 5{,}32$ soit maximal. La figure 8.2 représente les graphiques de la densité de probabilité $\varphi(y; a; \sigma^2)$ pour la valeur $\hat{a}_{mv} = \bar{y} = 5{,}243$ correspondant à la plus grande vraisemblance des observations $y_1 = 5{,}25$, $y_2 = 5{,}16$ et $y_3 = 5{,}32$ (la courbe pleine), ainsi que pour la valeur

$\hat{a} = 5{,}443$ pour laquelle les observations sont manifestement invraisemblables (la courbe en pointillé): dans les deux cas la variance σ^2 a été calculée à l'aide de l'estimation corrigée du maximum de vraisemblance et est égale à 0,0064).

Les bonnes propriétés des estimations obtenues par le maximum de vraisemblance plaident en faveur de cette approche. On peut prouver, en particulier, que si des conditions de régularité assez larges sont imposées à la loi $f(X; \Theta)$ (cf., par exemple, [8]), les estimations du maximum de vraisemblance $\hat{\Theta}_{\text{inv}}$ du paramètre Θ sont *convergentes, asymptotiquement sans biais, asymptotiquement normales et asymptotiquement efficaces* (c'est-à-dire que leur matrice asymptotique des covariances $\Sigma(\hat{\Theta}_{\text{inv}})$ est de la forme $\Sigma(\hat{\Theta}_{\text{inv}}) = n^{-1}\mathbf{I}^{-1}(\Theta; X)$, où $I(\Theta; X)$ est la matrice de Fisher définie par (8.7) dans le cas d'une seule observation, c'est-à-dire pour $\mathbf{X} = X$).

Mais cela ne signifie pas que les estimations du maximum de vraisemblance seront les meilleures dans tous les cas. Premièrement, leurs bonnes propriétés ne se manifestent souvent que dans des échantillons de taille élevée (c'est-à-dire sont asymptotiques, cf. § 8.4) de sorte que pour n petit elles sont concurrencées (et même surpassées) par d'autres estimations, par exemple, les estimations de la méthode des moments, de la méthode des moindres carrés, etc. (cf. plus bas n° 8.6.2 à 8.6.5). Deuxièmement, et ceci est le point faible de cette approche, pour construire des estimations par le maximum de vraisemblance et pour que leurs bonnes propriétés puissent jouer, *il est nécessaire de connaître exactement le type de loi de probabilité analysée*, ce qui est pratiquement impossible dans la plupart des cas. Dans ces conditions, il est plus payant de chercher non pas la meilleure estimation *de la densité* $f(X; \Theta)$ de la loi *étudiée* (estimation qui souvent perd ses bonnes propriétés dès que la densité réelle s'écarte de $f(X; \Theta)$), mais une estimation qui à défaut d'être la meilleure pour $f(X; \Theta)$ possède des propriétés *suffisamment stables* dans une plus vaste classe de densités comprenant $f(X; \Theta)$ comme cas particulier (cf. n° 8.6.4). De telles estimations sont dites *stables* ou *robustes*. Et, enfin, les estimations par le maximum de vraisemblance *peuvent même ne pas être convergentes* si le nombre k de paramètres $\theta_1, \ldots, \theta_k$ est élevé (est du même ordre de grandeur que la taille n de l'échantillon) et croît avec le nombre des observations. Un exemple d'une telle situation est exhibé plus bas (cf. exemple 8.7).

Essayons de répondre à la question de savoir comment déterminer concrètement les estimations par le maximum de vraisemblance, c'est-à-dire comment résoudre le problème d'optimisation (8.19).

Si la fonction $f(X; \Theta)$ satisfait certaines conditions de régularité (dérivabilité par rapport à Θ, etc., conditions a), b) et c) du § 8.3) et si l'extremum dans (8.19) est atteint en un point intérieur du domaine des valeurs possibles de Θ, alors les dérivées partielles de

la fonction $L(X_1, \ldots, X_n; \Theta)$ doivent s'annuler au point $\hat{\Theta}_{mv}$ et, par suite, celles du *logarithme de la fonction de vraisemblance*

$$l(X_1, \ldots, X_n; \Theta) = \ln L(X_1, \ldots, X_n; \Theta) = \\ = \sum_{i=1}^{n} \ln f(X_i, \Theta) \quad (8.20)$$

en raison de la monotonie de la relation liant L à l. Mais la fonction l est plus commode dans les calculs. Donc, dans le cas traité, l'estimation $\hat{\Theta}_{mv} = (\hat{\theta}^{(1)}_{mv}, \ldots, \hat{\theta}^{(k)}_{mv})$ doit satisfaire les équations

$$\frac{\partial l(X_1, \ldots, X_n; \Theta)}{\partial \theta^{(j)}} = 0, \; j = 1, 2, \ldots, k, \quad (8.21)$$

et peut être définie comme la solution de ce système d'équations.

Il existe des cas (correspondant à des lois de probabilité non régulières par rapport à Θ) où le système (8.21) n'est pas défini ou ne possède pas de solutions, alors que la solution de (8.19') existe. En pareille circonstance, il faut chercher l'estimation $\hat{\Theta}_{mv}$ par d'autres méthodes, y compris par une sélection directe de la solution de (8.19) (cf. plus bas, exemples 8.5 et 8.6).

Exemple 8.3. Considérons une variable aléatoire normale ξ de densité de probabilité

$$\varphi(x; a; \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-a)^3}{2\sigma^2}}$$

de moyenne $a = E\xi$ et de variance $\sigma^2 = V\xi$ inconnues.

En vertu de (8.5), la fonction de vraisemblance s'écrit

$$L(x_1, \ldots, x_n; a, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2}.$$

Le logarithme de la fonction de vraisemblance

$$l(x_1, \ldots, x_n; a, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\sigma^2 - \\ - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2.$$

En dérivant l par rapport à a et σ^2 et en annulant successivement les dérivées partielles, on obtient le système (8.21):

$$\frac{\partial l(x_1, \ldots, x_n; a, \sigma^2)}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - a) = 0;$$

$$\frac{\partial l(x_1, \ldots, x_n; a, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - a)^2 = 0.$$

La résolution de ce système par rapport à a et σ^2 nous donne les estimations par le maximum de vraisemblance de ces paramètres :

$$\hat{a}_{\text{mv}} = \frac{1}{n} \sum_{i=1}^{n} x_i \; ; \; \hat{\sigma}^2_{\text{mv}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

On a vu au § 8.3 que l'estimation $\hat{a}_{\text{mv}} = \bar{x}(n)$ est une estimation efficace du paramètre a (puisque son efficacité $e(\hat{a}_{\text{mv}}) = 1$). On en déduit que l'estimation $\hat{\sigma}^2_{\text{mv}} = s^2(n)$ est une estimation asymptotiquement efficace du paramètre σ^2 (son efficacité $e(\hat{\sigma}^2_{\text{mv}}) = \left(1 - \frac{1}{n}\right)^2 \Big/ \left(1 - \frac{1}{2n}\right)\Big)$.

Exemple 8.4. Considérons une variable aléatoire de Poisson ξ, de densité de probabilité

$$f(x; \lambda) = P\{\xi = x\} = \frac{\lambda^x}{x!} e^{-\lambda} \; (x = 0, 1, 2, \ldots),$$

où λ est un paramètre inconnu.

En vertu de (8.20), le logarithme de la fonction de vraisemblance construit sur l'échantillon $x_1, x_2, \ldots, x_n$ s'écrit

$$l(x_1, \ldots, x_n; \lambda) = \sum_{i=1}^{n} (x_i \ln \lambda - \ln(x_i!) - \lambda) =$$

$$= (\ln \lambda) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!) - n\lambda.$$

Une dérivation par rapport à λ nous conduit à l'équation de la méthode du maximum de vraisemblance

$$\frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0,$$

d'où

$$\hat{\lambda}_{\text{mv}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Il est immédiat que cette estimation est sans biais, puisque

$$\mathbf{E}\hat{\lambda}_{\text{mv}} = \mathbf{E}\left(\sum_{i=1}^{n} x_i\right)/n = \left(\sum_{i=1}^{n} \mathbf{E}x_i\right)/n = n\lambda/n = \lambda.$$

Calculons l'efficacité de $\hat{\lambda}_{\text{mv}}$. La borne inférieure de la variance sur toutes les estimations possibles du paramètre λ peut être calculée

à l'aide de l'inégalité d'information (8.12):

$$(V\hat{\lambda})_{\min}=\frac{1}{nE\left(\frac{d}{d\lambda}\ln f(x;\lambda)\right)^2}=\frac{1}{nE\left(\frac{x}{\lambda}-1\right)^2}=\frac{\lambda^2}{nE(x-\lambda)^2}=\frac{\lambda}{n}.$$

Calculons la variance de l'estimation $\hat{\lambda}_{mv}$ en partant du fait suivant (cf., par exemple, [40]): la somme de variables aléatoires indépendantes de Poisson $x_1, x_2, \ldots, x_n$ de moyennes respectives $\lambda_1, \lambda_2, \ldots, \lambda_n$ est une variable de Poisson de moyenne $\lambda_1+\lambda_2+\ldots$ $\ldots+\lambda_n$. Donc,

$$V\hat{\lambda}_{mv}=V\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)=\frac{1}{n^2}V\left(\sum_{i=1}^{n}x_i\right)=\frac{1}{n^2}\cdot n\lambda=\frac{\lambda}{n}.$$

En comparant $(V\hat{\lambda})_{\min}$ et $V\hat{\lambda}_{mv}$, on s'assure que l'estimation par le maximum de vraisemblance de la moyenne d'une variable aléatoire poissonnienne est efficace.

Exemple 8.5. Soit ξ une variable aléatoire uniforme (cf. n°6.1.7), c'est-à-dire telle que

$$f(x;\ a,\ b)=\begin{cases}\frac{1}{b-a} & \text{si } x\in[a,\ b];\\ 0 & \text{sinon,}\end{cases}$$

où a et b sont des paramètres inconnus (à estimer).

Il est immédiat de s'assurer que ce cas est irrégulier (ne serait-ce que parce que le domaine des valeurs possibles de ξ, dans lequel la densité est strictement positive, dépend des paramètres a et b). Donc la technique habituelle qui utilise les équations (8.21) de la méthode du maximum de vraisemblance est mise en défaut. Cependant, le problème d'extremum (8.19) peut être résolu *directement*.

En effet,

$$L(x_1,\ \ldots,\ x_n;\ a,\ b)=\frac{1}{(b-a)^n},$$

le domaine des valeurs possibles des paramètres a et b, dans lequel on cherche les valeurs $\hat{a}_{mv}$ et $\hat{b}_{mv}$ pour lesquelles $1/(b-a)^n=\max\limits_{a,\,b}$, étant défini par les relations

$$a\leqslant\min_{1\leqslant i\leqslant n}\{x_i\}=x_{\min}(n);$$

$$b\geqslant\max_{1\leqslant i\leqslant n}\{x_i\}=x_{\max}(n),$$

où $x_1, x_2, \ldots, x_n$ sont les valeurs observées de ξ. Il est évident que la solution du problème d'extremum

$$\max\frac{1}{(b-a)^n},\ a\leqslant x_{\min}(n),\ b\geqslant x_{\max}(n)$$

est

$$\hat{a}_{mv} = x_{min}(n); \quad \hat{b}_{mv} = x_{max}(n).$$

Les résultats du n° 5.6.4 (ainsi que du n° 8.6.4) nous permettent de calculer :

$$\begin{aligned} E\hat{a}_{mv} &= Ex_{min}(n) = a + \frac{b-a}{n+1}; \\ V\hat{a}_{mv} &= Vx_{min}(n) = \frac{n}{(n+1)^2(n+2)}(b-a)^2; \\ E\hat{b}_{mv} &= Ex_{max}(n) = b - \frac{b-a}{n+1}; \\ V\hat{b}_{mv} &= Vx_{max}(n) = \frac{n}{(n+1)^2(n+2)}(b-a)^2. \end{aligned} \tag{8.22}$$

On ne peut appliquer l'inégalité d'information au calcul de la borne inférieure de la variance de ces estimations, puisque ce cas est irrégulier. Sur (8.22) on voit que les quantités $V\hat{a}_{mv}$ et $V\hat{b}_{mv}$ caractérisent en même temps l'écart-type des estimations

$$\begin{aligned} \hat{a}'_{mv} &= x_{min}(n) - \frac{x_{max}(n) - x_{min}(n)}{n-1}; \\ b'_{mv} &= x_{max}(n) + \frac{x_{max}(n) - x_{min}(n)}{n-1} \end{aligned}$$

par rapport aux valeurs véritables des paramètres a et b.

Exemple 8.6. Revenons à l'estimation du paramètre de localisation θ dans une loi exponentielle de densité

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{pour } x \geqslant \theta; \\ 0 & \text{pour } x < \theta. \end{cases}$$

Comme dans l'exemple précédent, nous avons affaire à un cas *irrégulier*. Force est donc de résoudre directement le problème d'extremum

$$\begin{aligned} &\max_{\theta} L(x_1, \ldots, x_n; \theta) = \max_{\theta} e^{-\sum_{i=1}^{n}(x_i-\theta)}; \\ &\theta \leqslant \min_{1 \leqslant i \leqslant n} \{x_j\} = x_{min}(n). \end{aligned} \tag{8.22'}$$

Il est aisé de voir que $\hat{\theta}_{mv} = x_{min}(n)$ est solution de ce problème : pour tout autre θ vérifiant la condition (8.22′), il est évident que

$$\sum_{i=1}^{n}(x_i-\theta) = \sum_{i=1}^{n}[x_i-(\hat{\theta}_{mv}-\varepsilon)] = \sum_{i=1}^{n}(x_i-\hat{\theta}_{mv}) + \\ + n\varepsilon > \sum_{i=1}^{n}(x_i-\hat{\theta}_{mv}),$$

et, par suite,

$$L(x_1, \ldots, x_n ; \hat{\theta}_{mv}) > L(x_1, \ldots, x_n ; \theta).$$

Grâce aux résultats du n° 5.6.4 (ainsi que du n° 8.6.4), on peut calculer

$$\begin{aligned} &E\hat{\theta}_{mv} = Ex_{min}(n) = \theta + \frac{1}{n}; \\ &V\hat{\theta}_{mv} = E(x_{min}(n) - Ex_{min}(n))^2 = \frac{1}{n^2}; \\ &E(\hat{\theta}_{mv} - \theta)^2 = V\hat{\theta}_{mv} + \frac{1}{n^2} = \frac{2}{n^2}. \end{aligned} \tag{8.23}$$

Mais l'écart-type de l'estimation $\hat{\theta}'_{mv} = \hat{\theta}_{mv} - \frac{1}{n}$, déduite de l'estimation $\hat{\theta}_{mv}$ corrigée, vaut

$$E(\hat{\theta}'_{mv} - \theta)^2 = \frac{1}{n^2}. \tag{8.24}$$

Exemple 8.7 (emprunté à [138] p. 187). Etudions un cas où la méthode du maximum de vraisemblance ne fournit pas une estimation convergente.

Pour estimer n concentrations $a_1, a_2, \ldots, a_n$ d'un élément, on a procédé à une double mesure (x_i, y_i) de chaque concentration a_i. On admet que les $2n$ mesures $x_1, y_1, x_2, y_2, \ldots, x_n, y_n$ ont été effectuées avec la même précision et sont des variables aléatoires normales indépendantes (cf. n° 6.1.5), de sorte que la fonction de vraisemblance est

$$L(x_1, y_1, \ldots, x_n, y_n ; a_1, a_2, \ldots, a_n ; \sigma^2) = \\ = \frac{1}{(2\pi)^n \sigma^{2n}} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [(x_i - a_i)^2 + (y_i - a_i)^2]}.$$

Les paramètres inconnus sont les n moyennes $a_1, a_2, \ldots, a_n$ et la variance σ^2. On obtient sans peine l'estimation du maximum de vraisemblance des paramètres a_i :

$$\hat{a}_{i\,mv} = \frac{1}{2}(x_i + y_i).$$

La résolution de l'équation du maximum de vraisemblance (8.21), dans laquelle on aura substitué $\hat{a}_{i\,mv}$ à a_i, nous donne

$$\hat{\sigma}^2_{mv} = \frac{1}{4n} \sum_{i=1}^{n} (x_i - y_i)^2.$$

On établit sans peine que $E\hat{\sigma}^2_{mv} = \frac{\sigma^2}{2}$, c'est-à-dire que la méthode du maximum de vraisemblance nous conduit à une estimation de σ^2 de biais constant (asymptotiquement inévitable) égal à $-\frac{\sigma^2}{2}$. Pour meilleure estimation sans biais il aurait fallu prendre ici la statistique

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} (x_j - y_i)^2.$$

8.6.2. Méthode des moments. Soit ξ une variable aléatoire p-dimensionnelle dont la densité de probabilité $f(X; \Theta)$ si ξ est continue, ou la probabilité $P\{\xi = X \mid \Theta\}$ si ξ est discrète, dépend d'un paramètre $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ multidimensionnel. On se propose d'estimer le paramètre Θ, c'est-à-dire de construire une estimation $\hat{\Theta}$ sur le vu des observations indépendantes $X_1, \ldots, X_n$, où $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})'$.

La méthode des moments consiste à égaler un certain nombre de moments empiriques aux moments théoriques respectifs (c'est-à-dire calculés à l'aide de la fonction $f(X; \Theta)$) qui, à noter, sont manifestement fonctions des paramètres inconnus $\theta^{(1)}, \ldots, \theta^{(k)}$. En considérant un nombre de moments égal au nombre k des paramètres à estimer et en résolvant les équations obtenues par rapport à ces paramètres, on obtient les estimations cherchées. Donc, les estimations $\hat{\theta}^{(1)}_{mm}, \ldots, \hat{\theta}^{(k)}_{mm}$ de $\theta^{(1)}, \ldots, \theta^{(k)}$ par la méthode des moments sont solutions du système d'équations:

$$\begin{cases} \int x^{(l)} \cdot f(X; \Theta)\, dX = \frac{1}{n} \sum_{i=1}^{n} x_i^{(l)}, \quad l = 1, 2, \ldots, p; \\ \int x^{(l)} \cdot x^{(m)} \cdot f(X; \Theta)\, dX = \frac{1}{n} \sum_{i=1}^{n} x_i^{(l)} x_i^{(m)}, \quad l, m = 1, 2, \ldots p; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{8.25}$$

(si la variable aléatoire ξ est discrète, il faut remplacer les intégrales des premiers membres de (8.25) par les sommes correspondantes $\sum_j X_j^{0k} \cdot P\{\xi = X_j^0 \mid \Theta\}$).

Le nombre des équations du système (8.25) doit être égal au nombre k des paramètres à estimer. La question de savoir quels moments (initiaux, centrés ou certaines de leurs modifications du genre coefficients de dissymétrie ou d'aplatissement) il faut inclure dans le système (8.25) doit être résolue en fonction des objectifs concrets de l'analyse et en recherchant la plus simple forme de dépendance des caractéristiques *théoriques* alternatives des paramètres à estimer $\theta^{(1)}, \ldots, \theta^{(k)}$. Dans la pratique, on atteint rarement les moments

d'ordre quatre (exception faite de l'exploitation du « système de courbes de Pearson », cf. par exemple [27], mais cet appareil *purement formel* d'ajustage de la loi empirique sur la loi théorique est inapte, à notre sens, à résoudre des problèmes de statistique appliquée tant soi peu intéressants).

La méthode des moments a le mérite d'être assez facilement réalisable sur le plan numérique et de fournir des estimations (les solutions du système (8.25)) qui sont fonctions des moments empirique. Ceci simplifie l'étude des propriétés statistiques des estimations: on démontre (cf. [40], chap. 27 et 28) sous des conditions assez générales que ces estimations suivent une loi asymptotiquement normale pour de grands n, que leurs moyennes diffèrent des valeurs théoriques du paramètre d'une quantité de l'ordre de n^{-1} et que l'écart-type $\sigma(\hat{\theta}_{mm})$ est asymptotiquement de la forme $cn^{-1/2}$, où c est une constante.

Par ailleurs, Fisher (cf. [40]) a prouvé que l'efficacité asymptotique des estimations de la méthode des moments est généralement inférieure à l'unité, et de ce point de vue elles le cèdent aux estimations du maximum de vraisemblance. Mais ceci n'empêche pas la méthode des moments d'être très commode sur le plan pratique. Ses estimations sont prises parfois pour premières approximations et sont ensuite améliorées par des méthodes de plus grande efficacité.

Revenons à nos exemples.

Dans l'exemple 8.3, le système (8.25) s'écrit

$$\begin{cases} a = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i\,; \\ \sigma^2 + a^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2, \end{cases}$$

ce qui nous donne les estimations déjà obtenues par la méthode du maximum de vraisemblance:

$$\hat{a}_{mm} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}(n)\,;$$

$$\hat{\sigma}^2_{mm} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x}(n))^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}(n))^2 = s^2(n).$$

Les cas où l'échantillon suit la loi normale ou la loi de Poisson font partie des rares cas où les estimations de la méthode des moments sont confondues avec celles du maximum de vraisemblance (on s'en assure sans peine en s'adressant à l'exemple 8.4).

Dans l'exemple 8.5, le système (8.25) s'écrit :

$$\begin{cases} \dfrac{a+b}{2} = \bar{x}(n); \\ \dfrac{(b-a)^2}{12} = s^2(n). \end{cases}$$

D'où l'on déduit sans peine

$$\begin{aligned} \hat{a}_{\text{mm}} &= \bar{x}(n) - \sqrt{3}\, s(n); \\ \hat{b}_{\text{mm}} &= \bar{x}(n) + \sqrt{3}\, s(n). \end{aligned} \tag{8.26}$$

On peut comparer l'efficacité asymptotique des estimations acquises par la méthode du maximum de vraisemblance et par la méthode des moments : vu que la variance des estimations (8.26) en tant que variance de fonctions des moments empiriques $\bar{x}(n)$ et $s^2(n)$ est de l'ordre de n^{-1} (cf. [40]) et vu que la relation (8.22) qui exprime que la variance des estimations du maximum de vraisemblance des mêmes paramètres est de l'ordre de n^{-2}, on déduit que l'efficacité de $\hat{a}_{\text{mm}}$ et de $\hat{b}_{\text{mm}}$ par rapport à celle de $\hat{a}_{\text{mv}}$ et de $\hat{b}_{\text{mv}}$ tend vers 0 lorsque $n \to \infty$.

La réalisation de la méthode des moments dans l'exemple 8.6 nous donne

$$1 + \theta = \bar{x}(n).$$

Donc, $\hat{\theta}_{\text{mm}} = \bar{x}(n) - 1$.

Pour calculer la moyenne et la variance de l'estimation $\hat{\theta}_{\text{mm}}$, on se servira des faits suivants : a) la variable aléatoire x_i qui suit une loi exponentielle de paramètre $\lambda = 1$ et de paramètre de localisation θ (cf. n° 6.1.8) peut être interprétée comme un cas particulier d'une variable aléatoire répartie d'après une loi gamma de paramètres $a = 1$, $b = 1$ et de paramètre de localisation θ (cf. n° 6.2.5) ; b) la somme de n variables aléatoires indépendantes $x_1, x_2, \ldots, x_n$ suivant chacune une loi gamma de paramètres $a = 1$, $b = 1$ et de paramètre de localisation θ est répartie suivant une loi gamma de paramètres $a = n$, $b = 1$, et de paramètre de localisation θ (cf. n° 6.2.5). Donc,

$$\mathbf{E}\hat{\theta}_{\text{mm}} = \mathbf{E}(\bar{x}(n) - 1) = \frac{n + n\theta}{n} - 1 = \theta;$$

$$\mathbf{E}(\hat{\theta}_{\text{mm}} - \theta)^2 = \mathbf{V}\hat{\theta}_{\text{mm}} = \mathbf{V}(\bar{x}(n) - 1) = \frac{n}{n^2} = \frac{1}{n}.$$

En tenant compte de l'expression (8.24) de l'écart-type de l'erreur de l'estimation « corrigée » et $\hat{\theta}'_{\text{mv}}$ du paramètre θ, on obtient

$$\frac{\mathbf{E}(\hat{\theta}'_{\text{mv}} - \theta)^2}{\mathbf{E}(\hat{\theta}_{\text{mm}} - \theta)^2} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \to 0 \text{ lorsque } n \to \infty,$$

c'est-à-dire que dans ce cas l'efficacité asymptotique de l'estimation par la méthode des moments tend vers 0.

8.6.3. Méthode des moindres carrés. Considérons une fonction $\psi(\Theta, X)$ *de forme connue* dépendant d'un paramètre vectoriel inconnu $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ et d'une variable multidimensionnelle (non aléatoire) $X = (x^{(1)}, \ldots, x^{(p)})'$ caractérisant les conditions de réalisation de l'expérience aléatoire (ou de l'observation). Supposons qu'à l'issue de la i-ième expérience (observation) on enregistre la valeur y_i de la fonction $\psi(\Theta; X_i)$ avec une erreur ε_i (cf. aussi (3.9)) :

$$y_i = \psi(\Theta; X_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

sachant que la variable « auxiliaire » X_i est connue très exactement.

On demande d'estimer les paramètres $\theta^{(1)}, \ldots, \theta^{(k)}$ avec la meilleure précision possible sur le vu des observations $(y_1, X_1), \ldots, (y_n, X_n)$. *Contrairement aux procédures précédentes d'estimation* (cf. n° 8.6.1 et 8.6.2), *nous ne sommes pas dans l'obligation de nous donner la forme générale de la loi de probabilité des erreurs* ε_i (et partant des variables aléatoires y_i).

La *méthode des moindres carrés* définit l'estimation $\hat{\Theta}_{mc}$ de Θ à partir de la condition

$$\sum_{i=1}^{n} (y_i - \psi(\hat{\Theta}_{mc}; X_i))^2 = \min_{\hat{\Theta}} \sum_{i=1}^{n} (y_i - \psi(\hat{\Theta}, X_i))^2. \qquad (8.27)$$

Sous des conditions assez générales imposées à la nature des erreurs aléatoires ε et à la structure des fonctions $\psi(\Theta; X_i)$, les estimations satisfaisant la relation (8.27) sont convergentes, asymptotiquement sans biais, asymptotiquement normales et asymptotiquement efficaces (cf. par exemple [114], chap. 4). Indiquons quelques conditions essentielles que doivent remplir ε_i et $\psi(\Theta; X)$ pour que les estimations de la méthode des moindres carrés soient douées de bonnes propriétés :

a) les erreurs aléatoires ε_i ont des moyennes nulles ($\mathbf{E}\varepsilon_i = 0$) et des variances identiques finies $\mathbf{V}\varepsilon_i = \sigma^2$ ne dépendant ni du numéro i de l'observation ni du paramètre Θ *).

b) la fonction $\psi(\Theta; X)$ est continue et dérivable par rapport à tous les paramètres $\theta^{(1)}, \ldots, \theta^{(k)}$.

*) En fait, la condition essentielle est l'indépendance de la répartition des erreurs ε_i par rapport à Θ : si cette condition est violée, les estimations de la méthode des moindres carrés cessent d'être convergentes. Si l'on a affaire à une dépendance de la forme $\mathbf{V}\varepsilon_i = \sigma_i^2$, les estimations de la méthode des moindres carrés conservent leurs bonnes propriétés à condition d'introduire dans les termes de (8.27) des « poids » ω_i proportionnels à σ_i^{-2}.

La *méthode de calcul* des estimations de la méthode des moindres carrés repose sur le fait que si $\hat{\Theta}_{mc}$ est un point de minimum du critère

$$Q(\Theta) = \sum_{i=1}^{n} (y_i - \psi(\Theta, X_i))^2, \tag{8.27'}$$

les estimations $\hat{\theta}_{mc}^{(1)}, \ldots, \hat{\theta}_{mc}^{(k)}$ doivent satisfaire le *système d'équations normales*

$$\frac{\partial Q(\hat{\Theta}_{mc})}{\partial \hat{\theta}_{mc}^{(j)}} = 0, \quad j = 1, 2, \ldots, k.$$

Il est équivalent de dire que les estimations $\hat{\theta}_{mc}^{(1)}, \ldots, \hat{\theta}_{mc}^{(k)}$ des paramètres $\theta^{(1)}, \ldots, \theta^{(k)}$ sont solutions du système d'équations

$$\sum_{i=1}^{n} y_i \cdot \frac{\partial \psi(\hat{\Theta}_{mc}; X_i)}{\partial \hat{\theta}_{mc}^{(j)}} = \sum_{i=1}^{n} \psi(\hat{\Theta}_{mc}; X_i) \cdot \frac{\partial \psi(\hat{\Theta}_{mc}; X_i)}{\partial \hat{\theta}_{mc}^{(j)}},$$
$$j = 1, 2, \ldots, k. \tag{8.28}$$

Représentons les résultats décrits dans le cas particulier où la fonction ψ est *linéaire* par rapport aux variables X et aux paramètres Θ. En adoptant les notations matricielles du chapitre 3, plus exactement, en introduisant la matrice des observations

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(p)} \\ \cdot & \cdot & \cdot & \cdot \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p)} \end{pmatrix}$$

et les vecteurs colonnes des observations de la variable dépendante $Y = (y_1, y_2, \ldots, y_n)'$ et des erreurs aléatoires $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, on trouve (cf. (3.5)) que $Y = \mathbf{X}\Theta + \varepsilon$ *). De façon respective,

$$Q(\Theta) = (Y - \mathbf{X}\Theta)'(Y - \mathbf{X}\Theta), \tag{8.27''}$$

et le système d'équations normales devient

$$\mathbf{X}'\mathbf{X}\Theta = \mathbf{X}'Y. \tag{8.28'}$$

La forme matricielle de la solution de ce système est

$$\hat{\Theta}_{mc} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \tag{8.29}$$

*) Pour introduire des coefficients qui soient les termes constants de ces équations, il faut composer la première colonne de la matrice $\mathbf{X}$ avec des valeurs d'une variable « fictive » identiquement égales à l'unité.

Interprétation géométrique des estimations de la méthode des moindres carrés dans le cas linéaire. Considérons un espace vectoriel R^n à n dimensions et munissons-le de la distance $\rho(U, V)$ de deux vecteurs $U, V \in R^n$ en posant

$$\rho^2(U, V) = (U - V)'(U - V).$$

Considérons le sous-espace T de R^n engendré par les vecteurs colonnes de la matrice X ou, ce qui est équivalent, par tous les vecteurs de la forme $X\Theta$, où $\Theta \in R^p$. Il est évident que la dimension de T est confondue avec le rang r de X, donc elle est $\leqslant p$ et ne peut être égale à p que si $r = p$. Désignons par S l'ensemble des vecteurs de R^n perpendiculaires à T. La dimension de S est $n - r$. Tout vecteur U de R^n se décompose d'une seule façon en deux composantes perpendiculaires

$$U = U_T + U_S, \tag{8.30}$$

telles que $U_T \in T$ et $U_S \in S$. Ces composantes sont les projetées de U respectivement sur T et S.

L'estimation de Θ par la méthode des moindres carrés nous donne une valeur du vecteur $X\hat{\Theta} \in T$ telle que le vecteur $Y - X\hat{\Theta}$ soit de longueur minimale, ce qui exprime que chercher une estimation par la méthode des moindres carrés revient à projeter Y sur T, ceci étant, $Y - X\Theta \in S$. La décomposition de tout vecteur sous la forme (8.30) étant unique, le critère (8.27″) prend la même valeur pour toutes les estimations par la méthode des moindres carrés, ce qui a été déjà signalé plus haut.

Etudions maintenant plus en détail les projections de Y sur T et sur S. D'après l'hypothèse de base (3.6), le vecteur des erreurs suit une loi normale dans R^n, est de moyenne nulle et possède une variance égale à σ^2 dans toute direction. Mettons-le sous la forme $\varepsilon = \varepsilon_T + \varepsilon_S$. Alors

$$Y_T = X\Theta + \varepsilon_T; \tag{8.31}$$

$$Y_S = \varepsilon_S. \tag{8.32}$$

De (8.32), il s'ensuit immédiatement, compte tenu de la dimension de S et de la définition de χ^2 (cf. n° 6.2.1), que $Q(\hat{\Theta}_{mc})/\sigma^2$ suit une loi du $\chi^2(n - r)$. De là on peut proposer pour σ^2 l'estimation sans biais

$$\hat{\sigma}^2 = Q(\hat{\Theta}_{mc})/(n - r). \tag{8.33}$$

Optimalité des estimations de la méthode des moindres carrés. Lorsque $r = p$, la seule estimation de la méthode des moindres carrés est définie par la formule (8.29) qui, compte tenu de l'hypothèse (3.6), entraîne

$$E\hat{\Theta}_{mc} = \Theta, \tag{8.34}$$

c'est-à-dire que l'unique estimation de la méthode des moindres carrés est sans biais (cf. § 8.1). Montrons maintenant que parmi toutes les estimations linéaires sans biais de la forme $\hat{\Theta} = AY$ (telles que $AX\Theta = \Theta$) l'estimation $\hat{\Theta}_{mc}$ possède la plus petite variance généralisée (cf. n° 5.6.7) :

$$V_g\hat{\Theta}_{mc} = \sigma^2 \mid (X'X)^{-1} \mid. \tag{8.35}$$

A cet effet, projetons chaque vecteur ligne de la matrice A sur les sous-espaces T et S et formons avec ces projections les matrices respectives A_T et A_S. Comme $A = A_T + A_S$, il vient

$$\hat{\Theta} = AY = A_TY + A_SY; \tag{8.36}$$

$$E\hat{\Theta} = A_TX\Theta + A_SX\Theta. \tag{8.37}$$

Les vecteurs lignes de la matrice A_S appartiennent à S, c'est-à-dire sont perpendiculaires aux vecteurs colonnes X, donc le second terme de (8.37) est nul. L'estimation $\hat{\Theta}$ étant sans biais, il s'ensuit que les vecteurs $A_TX\Theta$ et Θ doivent être confondus pour toutes les valeurs de Θ. Ceci n'est possible, compte tenu du rang de X et du fait que les vecteurs lignes des matrices A_T et $(X'X)^{-1}X'$ appartiennent au sous-espace T, que si

$$A_T = (X'X)^{-1}X'. \tag{8.38}$$

Par ailleurs, en tenant compte de la décomposition (8.30), on trouve que

$$A_SY = A_SY_T + A_SY_S = A_SY_S, \tag{8.39}$$

puisque les vecteurs lignes A_S et Y_T appartiennent à des espaces orthogonaux. De (8.36), (8.38) et (8.39) il résulte que toute estimation sans biais linéaire de Θ peut être mise sous la forme

$$\hat{\Theta} = (X'X)^{-1}XY + A_SY_S = \hat{\Theta}_{mc} + A_SY_S; \tag{8.40}$$

signalons que les termes du second membre sont indépendants, car appartenant à des sous-espaces orthogonaux. L'optimalité de l'estimation de la méthode des moindres carrés découle directement de la représentation (8.40). En effet, la matrice des covariances des composantes de l'estimation de Θ s'écrit :

$$\begin{aligned} E(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)' &= E(\hat{\Theta}_{mc} - \Theta)(\hat{\Theta}_{mc} - \Theta)' + \\ &+ EA_SY_SY_S'A_S' = (X'X)^{-1}XEYY'X'(X'X)^{-1} + \\ &+ A_SEY_SY_S'A_S' = \sigma^2(X'X)^{-1} + \Sigma^*, \end{aligned}$$

où Σ^* est une matrice semi-définie positive.

Considérons quelques exemples particuliers.

1. Si les conditions de réalisation des observations ne varient pas, la fonction $\psi(\Theta, X)$ analysée ne dépendra pas de la variable auxiliaire X. Supposons, en particulier, que $\psi(\theta) = \theta$, de sorte que $y_i = \theta + \varepsilon_i$, c'est-à-dire que le problème revient à estimer un paramètre θ observé avec une erreur aléatoire dont la variance est éventuellement égale à $\sigma^2 = V\varepsilon_i$. Le critère de la méthode des moindres carrés devient dans ce cas: $Q(\theta) = \sum_{i=1}^{n} (y_i - \theta)^2$. Le système des équations normales (8.28) (qui est composé d'une seule équation) s'écrit

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\theta}_{mc},$$

d'où

$$\hat{\theta}_{mc} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}(n).$$

Si l'on admet de plus que l'erreur ε est normale, alors l'estimation $\hat{\theta}_{mc}$ est confondue avec l'estimation $\hat{\theta}_{mv}$ du maximum de vraisemblance de la moyenne d'une variable aléatoire normale.

2. Supposons que $\Theta = (\theta^{(1)}, \theta^{(2)})'$, $X = (x^{(1)}, x^{(2)})$, où $x^{(1)} \equiv 1$ (c'est-à-dire qu'il ne varie pas au cours des observations) et que $\psi(\Theta, X) = \Theta' \cdot X = \theta^{(1)} \cdot x^{(1)} + \theta^{(2)} \cdot x^{(2)} = \theta^{(1)} + \theta^{(2)} x^{(2)}$. Les observations sont

$$y_i = \theta^{(1)} + \theta^{(2)} x_i^{(2)} + \varepsilon_i.$$

On demande d'estimer les paramètres $\theta^{(1)}$ et $\theta^{(2)}$ sur le vu de ces observations (problème d'estimation des paramètres dans un modèle linéaire de régression double, cf. par exemple [7]).

Le critère de la méthode des moindres carrés s'écrit ici

$$Q(\theta^{(1)}, \theta^{(2)}) = \sum_{i=1}^{n} (y_i - \theta^{(1)} - \theta^{(2)} x_i^{(2)})^2.$$

Le système des équations normales (8.28′) est de la forme

$$\sum_{i=1}^{n} y_i = n \cdot \hat{\theta}_{mc}^{(1)} + \hat{\theta}_{mc}^{(2)} \sum_{i=1}^{n} x_i^{(2)};$$

$$\sum_{i=1}^{n} y_i x_i^{(2)} = \hat{\theta}_{mc}^{(1)} \sum_{i=1}^{n} x_i^{(2)} + \hat{\theta}_{mc}^{(2)} \sum_{i=1}^{n} (x_i^{(2)})^2,$$

d'où

$$\hat{\theta}_{\text{mc}}^{(2)} = \frac{\sum_{i=1}^{n} y_i x_i^{(2)} - n\bar{y}(n)\cdot\bar{x}^{(2)}(n)}{\sum_{i=1}^{n} (x_i^{(2)})^2 - n\,(\bar{x}^{(2)}(n))^2},$$

$$\hat{\theta}_{\text{mc}}^{(1)} = \bar{y}(n) - \hat{\theta}_{\text{mc}}^{(2)}\cdot\bar{x}^{(2)}(n),$$

où $\bar{y}(n)$ et $\bar{x}^{(2)}(n)$ sont comme toujours les moyennes arithmétiques respectivement des variables $y_1, \ldots, y_n$ et $x_1^{(2)}, \ldots, x_n^{(2)}$.

La méthode des moindres carrés est développée dans le détail, par exemple, dans [40], [114].

L'historique de la méthode des moindres carrés commence probablement en 1805, date à laquelle Legendre publia son travail « Nouvelles méthodes de définition des orbites des comètes », travail dans lequel fut proposée pour la première fois la fonctionnelle (8.27′) en tant que critère de qualité des estimations.

La première justification théorique de la méthode des moindres carrés fut donnée par Gauss en 1809 et en 1821. Dans sa forme générale, le théorème de Gauss relatif aux propriétés des estimations de la méthode des moindres carrés fut formulé et prouvé par Markov en 1912.

Deux circonstances ont contribué à la vaste diffusion de la méthode des moindres carrés dans les recherches statistiques: primo, elle n'implique pas la connaissance de la loi de probabilité des observations traitées, secundo, elle est suffisamment bien élaborée sur le plan de sa réalisation numérique.

8.6.4. Estimation par les statistiques « pondérées » ; censure, troncature des échantillons et statistiques de rang comme cas particulier de pondération. Les moments empiriques $m_k(n)$ sont toujours des estimations convergentes des moments théoriques respectifs m_k si ces derniers existent (cf. § 7.2). *Mais elles ne sont pas les plus efficaces pour toute population générale.* Nous avons vu par exemple (cf. n°s 8.6.1 et 8.6.2) que l'efficacité de l'estimation de la moyenne m_1 d'une variable à l'aide de la moyenne empirique $\bar{x}(n) = \hat{m}_1(n)$ dépendait essentiellement du type de la population générale analysée: pour une population générale normale, elle est égale à l'unité (cf. exemple 8.3 au n° 8.6.1); pour une population suivant une loi uniforme, elle est de beaucoup inférieure à l'unité et elle est d'ordre n^{-1} par rapport à l'efficacité de l'estimation

$$\hat{a}(n) = \frac{1}{2}\left(x_{\min}(n) + x_{\max}(n)\right) \tag{8.41}$$

(cf. exemple 8.5 au n°s 8.6.1 et 8.6.2). Pour construire l'estimation (8.41) nous n'avons utilisé que *deux* observations: la plus grande et

la plus petite, autrement dit, l'estimation (8.41) fait partie des statistiques de rang « pondérées » $\sum_{i=1}^{n} \omega_i x_{(i)}$, où $x_{(i)}$ est la i-ième observation par ordre de grandeur croissante, ω_i son « poids » (il est évident que dans la statistique (8.41) on a pris $\omega_1 = \omega_n = 0{,}5$ et tous les autres $\omega_i = 0$).

En réalité, la loi étudiée peut ne correspondre à aucun des modèles de lois décrites, par exemple, au chapitre 6, et être « à cheval » sur deux lois. Un procédé de description formelle d'une telle situation a été proposé par J. Tukey (cf. [136] et n° 6.1.11) pour les lois symétriques (par rapport au paramètre de groupement θ):

$$f(x;\ a;\ \varepsilon) = (1 - \varepsilon)\,\varphi(x;\ a;\ \sigma^2) + \varepsilon \cdot h(x;\ a;\ \theta),$$

où $f(x;\ a;\ \varepsilon)$ est la densité de probabilité de la variable étudiée; $\varphi(x;\ a;\ \sigma^2)$ la densité d'une loi normale de moyenne a et de variance σ^2; $h(x;\ a;\ \theta)$ la densité d'une loi (« polluante ») symétrique par rapport au même centre de groupement et dépendant, éventuellement, d'un paramètre θ; $\varepsilon \geqslant 0$ le taux de « pollution », c'est-à-dire le pourcentage des observations de la population générale étudiée qui suivent la loi de probabilité « polluante » $h(x;\ a;\ \theta)$.

Le tableau 8.1 représente les valeurs des efficacités asymptotiques des estimations $\bar{x}(n)$ (moyenne empirique) et $x_{\text{méd}}(n)$ (médiane empirique) du paramètre a en fonction du caractère et du taux de pollution de la loi analysée. Les calculs ont été effectués en combinant la méthode analytique qui utilise la connaissance de la loi de probabilité de $x_{\text{méd}}(n)$ dans les diverses populations générales (cf. n° 5.6.4) et la méthode de simulation statistique sur ordinateur (cf. § 6.3) sur des échantillons d'assez grandes tailles ($n \geqslant 200$).

Nous remarquons que plus les ailes de la loi polluante sont « lourdes » (c'est-à-dire plus la densité $h(x)$ tend lentement vers 0 à mesure que x s'éloigne du centre de groupement a) et plus l'efficacité de l'estimation $\bar{x}(n)$ faiblit, tandis que la médiane empirique $x_{\text{méd}}(n)$ affiche une remarquable stabilité de ses bonnes propriétés. La médiane empirique $x_{\text{méd}}(n)$ se rapporte à la classe des statistiques de rang « pondérées », c'est-à-dire des statistiques de la forme $\sum_{i=1}^{n} \omega_i\, x_{(i)}$; pour l'obtenir comme cas particulier des statistiques de cette classe, il suffit d'égaler à zéro tous les poids ω_i à l'exception d'un ($\omega_{\frac{n+1}{2}} = 1$ si n est impair) ou de deux $\left(\omega_{\frac{n}{2}} = \omega_{\frac{n}{2}+1} = \frac{1}{2}\right.$ si n est pair$\left.\right)$.

Par ailleurs, aux n°s 10.4.4, 10.4.5 et 10.4.6, ainsi qu'au § 11.5, on résout certains problèmes concrets d'estimation statistique de

Tableau 8.1

n°	Forme de pollution	Estimation	Taux de pollution		
			$\varepsilon = 0$	$\varepsilon = 0{,}05$	$\varepsilon = 1$
1	$h(x) = \frac{1}{\sqrt{2\pi}\,3\sigma} \times e^{\frac{-(x-\theta)^2}{2\cdot(3\sigma)^2}}$ (loi normale)	$\bar{x}(n)$	1	0,808	1
		$x_{\text{méd}}(n)$	$\frac{2}{\pi} = 0{,}637$	0,669	$\frac{2}{\pi} = 0{,}637$
2	$h(x) = \frac{1}{2} e^{-\lvert x-\theta \rvert}$ (loi de Laplace)	$\bar{x}(n)$	1	0,682	0,500
		$x_{\text{méd}}(n)$	$\frac{2}{\pi} = 0{,}637$	0,894	1
3	$h(\bar{x}) = \begin{cases} \frac{1}{6\cdot\sigma} \text{ si } x \in \\ \in [a-3\sigma,\ a+3\sigma]; \\ 0 \text{ si } x \notin [a-3\sigma, \\ a+3\sigma] \end{cases}$ (loi uniforme sur $[a-3\sigma,\ a+3\sigma]$)	$\bar{x}(n)$	1	0,362	0
		$x_{\text{méd}}(n)$	$\frac{2}{\pi} = 0{,}637$	0,781	0,835
4	$h(x) = \frac{1}{\pi[1+(x-a)^2]}$ (loi de Cauchy)	$\bar{x}(n)$	1	0	0
		$x_{\text{méd}}(n)$	$\frac{2}{\pi} = 0{,}637$	0,765	$\frac{8}{\pi^2} = 0{,}811$

paramètres inconnus à l'aide des statistiques pondérées. Arrêtons-nous succinctement sur la description des principales approches faisant intervenir les statistiques pondérées et sur la classification de leurs types.

Pondération des données empiriques $X_1, \ldots, X_n$. Dans le cas général, on attribue à l'observation X_i un poids $\omega_i = \omega(X_i) \geqslant 0$ qui est défini comme une fonction de sa valeur courante. Les poids sont généralement soumis à la condition de normalisation $\sum_{i=1}^{n} \omega(X_i) \equiv \equiv 1$. En particulier, on peut traiter les moments pondérés de la variable aléatoire ξ de densité $f_\xi(X)$ comme les moments aussi bien

empiriques $\hat{m}_k(n, \omega)$ que théoriques $\hat{m}_k(\omega)$ *):

$$\hat{m}_k(n, \omega) = \sum_{i=1}^{n} X_i^{(k)} \cdot \omega(X_i);$$

$$m_k(\omega) = \int X^k \omega(X) f_\xi(X)\, dX.$$

Les possibilités et le contenu de cette approche sont étudiés au n° 10.4.6.

Si l'on a affaire à des observations d'une variable à une dimension $x_1, \ldots, x_n$, le poids de chacune d'elles est défini en fonction de son numéro d'ordre dans une série d'observations ordonnées (par ordre de croissance), c'est-à-dire qu'on forme une *série variationnelle* $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ (cf. n° 5.6.4) et à chaque terme $x_{(i)}$ on associe un poids, ω_i.

On a déjà cité des exemples d'une telle pondération (qui nous a conduits aux statistiques de rang). D'autres variantes d'estimations de ce type sont envisagées aux n°s 10.4.4 et 10.4.5.

Censure des échantillons. Cette méthode consiste à attribuer des *poids nuls* aux termes extrêmes de la série variationnelle et des poids identiques strictement positifs aux autres. Si les poids nuls sont attribués aux seules observations extérieures à un intervalle donné $[a, b]$, c'est-à-dire que

$$\omega(x_{(i)}) = \begin{cases} \omega_0 > 0 & \text{si } x_{(i)} \in [a, b]; \\ 0 & \text{si } x_{(i)} \notin [a, b], \end{cases}$$

on parle alors d'une *censure de type I*. Il est évident que dans ce cas le nombre ν des autres observations est une quantité *aléatoire* ($\nu < n$).

Si les poids nuls sont attribués à un pourcentage fixe α des plus petites observations et à un *pourcentage* fixe β des plus grandes, on parle alors d'une *censure de type II de niveau* (α, β). Dans ce cas, le nombre ν des autres observations est connu à l'avance et vaut $n(1 - \alpha - \beta)$.

Le chercheur peut être amené à censurer *de gré* ou *de force*. La censure forcée est dictée par les conditions de l'expérience: si, par exemple, nous testons n articles à la destruction dans un intervalle de temps limité T, il est évident que nous aurons affaire à une censure de type I qui exclue les valeurs exactes de la durée de vie (temps précédant la destruction) de toutes les pièces qui n'auront pas été détruites durant l'intervalle T. Par ailleurs, dans la classe des esti-

*) Si $X = (x^{(1)}, \ldots, x^{(p)})'$ est une variable multidimensionnelle, par X^k on comprend tout produit de la forme $(x^{(i_1)})^{k_1}, (x^{(i_2)})^{k_2} \ldots (x^{(i_t)})^{k_t}$, dans lequel les indices i_l peuvent prendre toutes les valeurs entre 1 et p, et $k_1 + k_2 + \ldots + k_t = k$.

mations construites sur les échantillons censurés, on peut souvent trouver des estimations qui à défaut d'être les meilleures dans le cadre rigide d'une population générale d'un type donné sont caractérisées par une stabilité appréciable de leurs bonnes propriétés pour les écarts par rapport aux hypothèses de base (cf. plus haut l'exemple relatif à la comparaison entre la moyenne empirique et la médiane).

Troncature d'une loi. Cette situation se présente lorsque la variable aléatoire ξ ne peut être observée sur une partie de son domaine de valeurs possibles. Si, par exemple, on étudie la répartition des foyers en fonction du revenu et que les conditions de réalisation du sondage nous privent de la possibilité d'analyser les éléments dont le revenu est inférieur à un niveau donné a (en roubles), on dit alors que la *loi est tronquée à gauche au point a*. Contrairement aux échantillons censurés, nous n'avons même pas la possibilité d'estimer ici le pourcentage d'observations situées au delà du niveau de troncature.

Dans [77], chap. 32 et [41] par exemple, le lecteur trouvera des renseignements très détaillés sur l'utilisation, dans les problèmes d'estimation statistique des paramètres, des statistiques pondérées et, en particulier, de rang, et des statistiques construites sur les échantillons censurés, ainsi que la discussion des divers problèmes de stabilité des estimations obtenues.

8.6.5. Construction des intervalles de confiance (régions de confiance). Au § 8.5 nous avons introduit la notion d'intervalle de confiance du paramètre inconnu $\Theta = (\theta^{(1)}, \ldots, \theta^{(k)})'$ et de région de confiance (pour $k \geqslant 2$). Comment construire à l'aide des données empiriques $X_1, \ldots, X_n$ un domaine aléatoire $\Delta\Theta_P(X_1, \ldots, X_n)$ qui recouvre la valeur inconnue du paramètre Θ avec un niveau de confiance P donné *a priori*? Il est évident que ce domaine doit être construit autour d'une estimation *ponctuelle* $\hat{\Theta}$ de Θ; quant à sa forme exacte et son volume, ils dépendent de la loi de probabilité de la variable aléatoire $\hat{\Theta}$ et, en particulier, de sa fonction de répartition $F_{\hat{\Theta}}(U; \Theta)$ qui malheureusement dépend aussi de la valeur véritable inconnue du paramètre Θ.

Il existe deux méthodes pour tourner cette difficulté. La *première* consiste, si cela est possible, à construire des régions de confiance *exactes* $\Delta\Theta_P(X_1, \ldots, X_n)$ (pour toute taille finie n de l'échantillon) et à trouver des fonctions ψ_1 et ψ_2 de k variables et des constantes de normalisation $A(X_1, \ldots, X_n)$ et $B(X_1, \ldots, X_n)$ ne dépendant pas de Θ, telles que la loi des statistiques

$$\gamma_1(X_1, \ldots, X_n) = A(X_1, \ldots, X_n) \cdot \psi_1(\hat{\Theta} - \Theta)$$

ou

$$\gamma_2(X_1, \ldots, X_n) = B(X_1, \ldots, X_n) \cdot \psi_2\left(\frac{\hat{\theta}^{(1)}}{\theta^{(1)}}, \ldots, \frac{\hat{\theta}^{(k)}}{\theta^{(k)}}\right)$$

puisse être décrite exactement (par exemple, à l'aide d'une loi classique tabulée, cf. n°s 6.1.5, 6.2.1, 6.2.6) et *ne dépende pas du paramètre inconnu* Θ.

Déterminons à titre d'exemple l'intervalle de confiance des paramètres a et σ^2 d'une population générale normale (cf. exemple 8.3 au n° 8.6.1).

On sait (cf. n° 6.2.2) que la statistique

$$\frac{(\bar{x}(n)-a)\sqrt{n-1}}{s(n)}$$

suit une loi de Student à $n-1$ degrés de liberté (ici la fonction $\psi_1(\hat{\theta}-\theta)=\hat{\theta}-\theta$, et la constante de normalisation $A(x_1, \ldots, x_n)=\sqrt{n-1}/s(n)$). Donc, après avoir défini tabulairement les points de pourcentage d'ordre $q=(1-P)/2$ et $1-q=(1+P)/2$ de la loi t de Student à $n-1$ degrés de liberté correspondant à la probabilité P donnée (c'est-à-dire les points $w_{100q}(t(n-1))$ et $w_{100(1-q)}(t(n-1))$, on peut affirmer, compte tenu de la symétrie de cette loi en vertu de laquelle $w_{100(1-q)}=-w_{100q}$ (cf. n° 5.6.5), que l'inégalité

$$-w_{100q}(t(n-1))<\frac{(\bar{x}(n)-a)\sqrt{n-1}}{s(n)}<w_{100q}(t(n-1))$$

est réalisée avec une probabilité $P=1-2q$. Or ceci exprime que l'intervalle de confiance aléatoire

$$\Delta a_P(x_1, \ldots, x_n)=$$
$$=\left[\bar{x}(n)-w_{100q}(t(n-1))\cdot\frac{s(n)}{\sqrt{n-1}},\ \bar{x}(n)+w_{100q}(t(n-1))\cdot\frac{s(n)}{\sqrt{n-1}}\right]$$

recouvre la moyenne inconnue a avec la probabilité donnée P.

Pour construire l'intervalle de confiance du paramètre σ^2, on se servira du fait que la statistique $\frac{ns^2(n)}{\sigma^2}$ suit une loi de χ^2 à $n-1$ degrés de liberté (cf. n° 6.2.1). Donc, ici $\psi_2(\hat{\theta}/\theta)=\hat{\theta}/\theta$, et la constante de normalisation $B(x_1, \ldots, x_n)=n$. Après avoir défini tabulairement les points de pourcentage $w_{1-q}(\chi^2(n-1))$ et $w_q(\chi^2(n-1))$, où comme précédemment $q=(1-P)/2$ et P est le niveau de confiance donné, on peut affirmer que l'inégalité

$$w_{1-q}(\chi^2(n-1))<\frac{ns^2(n)}{\sigma^2}<w_q(\chi^2(n-1))$$

est réalisée avec la probabilité $P=1-2q$. Ce qui exprime que l'intervalle de confiance aléatoire

$$\Delta\sigma^2_P(x_1, \ldots, x_n)=\left[s^2(n)\cdot\frac{n}{w_{1-q}(\chi^2(n-1))},\ s^2(n)\cdot\frac{n}{w_q(\chi^2(n-1))}\right]$$

recouvre la valeur inconnue de la variance σ^2 avec la probabilité P.

La *deuxième méthode* de construction des régions de confiance, plus simple et plus universelle, est basée sur les propriétés asymptotiques des estimations et de ce fait fournit des résultats approchés qui ne la rendent utilisable que pour les échantillons d'assez grande taille. Cette méthode utilise le fait (cf. § 8.4) que les estimations du maximum de vraisemblance et les estimations de la méthode des moments suivent une loi conjointe asymptotiquement normale, c'est-à-dire que la loi du vecteur à k dimensions $\xi(n) = \sqrt{n}\,(\hat{\Theta} - \Theta)$ tend vers une loi normale multidimensionnelle de vecteur des moyennes nul et de matrice des covariances $\Sigma(\Theta)$ dépendant du paramètre inconnu Θ. Cette méthode « pèche » sur deux points: premièrement, la forme *asymptotique* de la loi de la variable aléatoire $\xi(n)$ est utilisée pour les échantillons de taille n *finie*; deuxièmement, l'estimation $\hat{\Theta}$ est substituée à la valeur inconnue de Θ dans la matrice $\Sigma(\Theta)$.

Pour construire la région de confiance du paramètre inconnu $\boldsymbol{\Theta} = (\theta^{(1)}, \ldots, \theta^{(k)})'$, nous devons maintenant nous servir du fait notoire suivant (cf. [16]): si le vecteur à k dimensions $\xi(n) = \sqrt{n}\,(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})$ suit une loi normale de paramètrres $\mathbf{0}$ et $\Sigma(\boldsymbol{\Theta})$, la variable aléatoire

$$n(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})' \cdot \Sigma^{-1}(\theta)(\boldsymbol{\Theta} - \theta)$$

est répartie suivant une loi du χ^2 à k degrés de liberté.

Après avoir défini tabulairement les points de pourcentage $w_{1-q}(\chi^2(k))$ et $w_q(\chi^2(k))$, où $q = (1 - P)/2$, correspondant à la probabilité P, et après avoir substitué l'estimation $\hat{\boldsymbol{\Theta}}$ à la valeur inconnue de $\boldsymbol{\Theta}$ dans la matrice *connue* $\Sigma(\boldsymbol{\Theta})$, on peut affirmer que l'inégalité

$$w_{1-q}(\chi^2(k)) < n(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})' \cdot \Sigma^{-1}(\hat{\boldsymbol{\Theta}}) \cdot (\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) < w_q(\chi^2(k)) \qquad (8.42)$$

est réalisée avec une probabilité θ approximativement égale à P.

Remarque 1. Si l'on a un seul paramètre θ à estimer (c'est-à-dire si $k = 1$), on peut se servir directement du fait que la différence $\hat{\theta} - \theta$ suit une loi normale de paramètres $(0, \sigma^2_{\hat{\theta}}(\theta))$ et au lieu de (8.42) écrire

$$-w_q(\xi(0, 1))\,\sigma_{\hat{\theta}}(\theta) < \hat{\theta} - \theta < w_q(\xi(0, 1)) \cdot \sigma_{\hat{\theta}}(\theta), \qquad (8.42')$$

où $w_q(\xi(0, 1))$ est un point de pourcentage $100q$ de la loi normale réduite, $\sigma^2_{\hat{\theta}}(\theta)$ la variance de l'estimation $\hat{\theta}$. De (8.42′) on déduit l'intervalle de confiance

$$\Delta\theta_P(x_1, \ldots, x_n) = [\hat{\theta} - w_q(\xi(0, 1)) \cdot \sigma_{\hat{\theta}}(\theta),\ \hat{\theta} + w_q(\xi(0, 1)) \times \sigma_{\hat{\theta}}(\theta)]. \qquad (8.42'')$$

Remarque 2. Si pour $\hat{\Theta}$ on prend des estimations ponctuelles du maximum de vraisemblance, la matrice des covariances $\Sigma(\Theta)$ du vecteur $\sqrt{n}(\hat{\Theta}-\Theta)$ est définie sans ambiguïté par la matrice de Fisher (cf. §§ 8.3 et 8.4):

$$\Sigma(\Theta) = \mathrm{I}^{-1}(\Theta),$$

où les éléments de la matrice $\mathrm{I}(\Theta)$ sont définis par les relations (8.7).

Remarque 3. La définition positive et la symétrie de la matrice Σ conditionnent l'*ellipsoïdalité* de la région de confiance définie par la relation (8.42).

Exemple 8.8. Soit à définir sur le vu des observations $x_1, \ldots, x_n$ l'intervalle de confiance du paramètre p de la loi binomiale (cf. n° 6.1.1), c'est-à-dire de la loi d'une variable aléatoire discrète ξ de densité

$$f(x;\, p,\, N) = P\{\xi = x\} = C_N^x p^x (1-p)^{N-x}, \quad x = 0,\, 1,\, \ldots,\, N,$$

où $N > 0$ est un entier connu, $p \in]0,\, 1[$ le paramètre à estimer.

Commençons, en vertu de la technique décrite au n° 8.6.1, par calculer l'estimation ponctuelle $\hat{p}$ de p par le maximum de vraisemblance.

Le logarithme de la fonction de vraisemblance s'écrit ici

$$l(x_1,\, x_2,\, \ldots,\, x_n;\, p,\, N) = \\ = \sum_{i=1}^{n} \ln C_N^{x_i} + \ln p \sum_{i=1}^{n} x_i + \ln(1-p) \sum_{i=1}^{n} (N - x_i).$$

L'équation respective du maximum de vraisemblance est

$$\frac{dl}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (N - x_i) = 0.$$

En la résolvant par rapport à p, on obtient l'estimation du maximum de vraisemblance

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{nN} = \frac{\bar{x}(n)}{N}.$$

L'indépendance des x_i et le fait que $\mathrm{E}x_i = Np$ et $\mathrm{V}x_i = Np(1-p)$ (cf. n° 6.1.1) nous donnent

$$\mathrm{E}\hat{p} = p; \quad \mathrm{V}\hat{p} = \frac{p(1-p)}{nN}.$$

En se donnant le niveau de confiance $P = 0{,}95$, en utilisant le fait que $\hat{p} - p$ suit une loi asymptotiquement normale et en substi-

tuant $\hat{p}$ à p dans l'expression de la variance $V\hat{p}$, on obtient, en vertu de (8.42″), l'intervalle de confiance

$$\Delta p_{0.95}(x_1, \ldots, x_n) = \left[\hat{p} - 1{,}96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{nN}},\ \hat{p} + 1{,}96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{nN}}\right].$$

8.6.6. Approche bayesienne de l'estimation statistique. L'idée maîtresse de l'approche bayesienne consiste à utiliser dans l'estimation d'un paramètre θ (qui peut éventuellement être vectoriel), en plus de l'information fournie par l'échantillon X, une information supplémentaire *a priori* sur le paramètre à estimer. On admet que le paramètre θ est une variable aléatoire possédant une *densité de probabilité a priori* $h_0(\theta)$ connue du chercheur.

La formule de Bayes (cf. n°s 4.1.3 et 4.2.2) nous permet de calculer la *densité de probabilité a posteriori* $h(\theta/X)$ du paramètre θ après les observations de l'échantillon X:

$$h(\theta \mid X) = \frac{f(X \mid \theta)}{g(X)} h_0(\theta),$$

où $f(X \mid \theta)$ est la densité de la loi conditionnelle de X sachant θ, c'est-à-dire la fonction de vraisemblance introduite au § 8.2; $g(X) = \int f(X \mid \theta)\, h_0(\theta)\, d\theta$ une constante de normalisation indépendante de θ.

En se servant de la densité $h(\theta \mid X)$, on peut construire par exemple l'*intervalle de confiance bayesien* du paramètre θ.

L'estimation bayesienne ponctuelle de θ *est la moyenne* calculée à l'aide de la loi *a posteriori*:

$$\hat{\theta}_{B_1}(X) = E_\theta[\theta \mid X] = \int \theta h(\theta \mid X)\, d\theta. \tag{8.43}$$

Signalons une importante propriété de l'estimation $\hat{\theta}_{B_1}$. Soit $t(X)$ une estimation dépendant de l'échantillon X. On appelle *risque bayesien a posteriori* la quantité

$$M_t(X) = E_\theta[(\theta - t(X))^2 \mid X] = \int (\theta - t(X))^2 h(\theta \mid X)\, d\theta. \tag{8.44}$$

L'estimation $\hat{\theta}_{B_1}$ en tant qu'espérance mathématique de la répartition conditionnelle de θ pour X donné minimise $M_t(X)$, et la valeur de $M_t(X)$ est simplement la variance de la répartition *a posteriori* *). Des propriétés de l'espérance mathématique conditionnelle (cf. n° 5.6.7) il s'ensuit que $\hat{\theta}_{B_1}$ minimise le *carré moyen complet de*

*) Une approche plus générale définit les estimations bayesiennes à partir de la condition de minimum d'intégrale de type (8.44), mais avec une fonction de perte $w(|\theta - t(X)|)$ qui est susceptible d'être différente de la fonction quadratique, cf. [141].

l'erreur, c'est-à-dire que

$$\hat{\theta}_{B_1} = \operatorname{Arg}\min_t M_t,$$

où

$$\left.\begin{aligned} M_t &= \mathbf{E}_{\mathbf{X}}\,[M_t\,(\mathbf{X})] = \mathbf{E}_\theta\,[M_t\,(\theta)]\,; \\ M_t\,(\theta) &= \mathbf{E}_{\mathbf{X}}\,[(\theta - t\,(\mathbf{X}))^2 \mid \theta], \end{aligned}\right\} \tag{8.45}$$

et l'erreur M_t (X) est définie dans (8.44).

On obtient l'autre estimation bayesienne $\hat{\theta}_{B_2}$ en choisissant la valeur de θ qui maximise la densité conditionnelle *a posteriori* $h\,(\theta \mid \mathbf{X})$:

$$\hat{\theta}_{B_2} = \operatorname{Arg}\max_\theta h\,(\theta \mid \mathbf{X}), \tag{8.46}$$

c'est-à-dire l'estimation du maximum de vraisemblance « corrigée » par la présence de la densité *a priori* h_0 (θ). Si f (X | θ) et h_0 (θ) sont assujetties à des conditions assez faibles, les deux estimations $\hat{\theta}_{B_1}$ et $\hat{\theta}_{B_2}$ convergent lorsque $n \to \infty$ vers l'estimation du maximum de vraisemblance *indépendamment du choix de la densité a priori* h_0 (θ).

Exemple 8.9. Soit à estimer la probabilité inconnue p dans une suite d'épreuves de Bernoulli sur le vu des résultats de n épreuves indépendantes. Pour loi *a priori* de p, prenons une loi β (cf. n° 6.2.6) de densité

$$h_0\,(p) = \frac{p^{a-1}\,(1-p)^{b-1}}{B\,(a,\ b)}, \quad 0 < p < 1.$$

Supposons que n épreuves aient donné lieu à r succès. La densité *a posteriori* est alors

$$h\,(p/r) = \frac{p^{a+r-1}\,(1-p)^{b+n-r-1}}{B\,(a+r,\ b+n-r)},$$

c'est-à-dire est la densité d'une loi β avec d'autres paramètres. En prenant l'espérance mathématique de la loi de densité $h\,(p/r)$, on obtient l'estimation bayesienne

$$\hat{p}_{B_1} = \frac{a+r}{a+b+n},$$

qui diffère de l'estimation ordinaire r/n.

Exemple 8.10. Etudions le cas d'un paramètre scalaire discret θ. Le passage au cas continu décrit plus haut n'apporte de toute évidence aucune complication. Supposons, pour fixer les idées, que le paramètre θ ne peut prendre que deux valeurs: $\theta_1 = 0$ et $\theta_2 = 1$. La répartition *a priori* est définie par les probabilités *a priori* q_i pour les valeurs θ_i ($i = 1, 2$). Supposons qu'on ait réalisé une observation X (c'est-à-dire qu'on ait obtenu un échantillon composé d'une seule

observation). La répartition *a posteriori* du paramètre θ sera alors

$$h(\theta=0 \mid X)=\frac{q_1 f(X \mid \theta=0)}{q_1 f(X \mid \theta=0)+q_2 f(X \mid \theta=1)};$$
$$h(\theta=1 \mid X)=1-h(\theta=0 \mid X).$$

En se servant de (8.46), on obtient l'estimation bayesienne

$$\hat{\theta}_{B_2}=\begin{cases} 1 & \text{si} \quad h(\theta=1/X)>h(\theta=0 \mid X); \\ 0 & \text{si} \quad h(\theta=1/X)\leqslant h(\theta=0 \mid X). \end{cases} \tag{8.47}$$

En fait, nous avons obtenu une *règle de classification* de l'observation X dans l'un ou l'autre de deux ensembles de fonctions de répartition *a priori* $\varphi_1(X)=f(X \mid \theta=0)$ et $\varphi_2(X)=f(X \mid \theta=1)$ appelée *règle de décision par le maximum de la probabilité a posteriori.* La règle (8.47) est équivalente à la règle suivante:

$$\hat{\theta}_{B_2}=\begin{cases} 1 & \text{si} \quad f(X \mid 1)/f(X \mid 0)>q_2/q_1; \\ 0 & \text{si} \quad f(X \mid 1)/f(X \mid 0)\leqslant q_2/q_1. \end{cases}$$

La quantité $f(X \mid 1)/f(X \mid 0)$ s'appelle *rapport de vraisemblance* (cf. § 9.3).

L'estimation $\hat{\theta}_{B_2}$ (la règle de classification par le maximum de la probabilité *a posteriori*) est optimale en ce sens qu'elle minimise la probabilité moyenne de l'erreur de classification

$$P_{ec}=q_1 P(1 \mid 0)+q_2 P(0 \mid 1),$$

où $P(1 \mid 0)$ est la probabilité d'estimer à tort le paramètre θ égal à 1, alors qu'en réalité il est égal à 0 (la probabilité $P(0 \mid 1)$ s'interprète de la même manière *mutatis mutandis*).

Le champ d'application de l'estimation bayesienne est limité par le fait que la justification du choix de la loi *a priori* est très compliquée et n'admet probablement pas de solution satisfaisante (cf. [39]).

Dans bien des cas, on peut toutefois disposer d'une certaine information sur la forme de la loi *a priori* à quelques paramètres inconnus près qui peuvent être estimés sur le vu de l'échantillon en même temps que le paramètre θ. Cette approche s'appelle *approche bayesienne empirique.*

Considérons maintenant une importante classe d'estimations dites *estimations du minimax* qui permettent de se passer de toute hypothèse sur la répartition *a priori.* Soit $t(X)$ une estimation de θ. Si la répartition *a priori* de θ est connue (par exemple la densité $h_0(\theta)$ est donnée), le carré moyen complet de l'erreur M_t est défini par (8.45). De la relation (8.45) il s'ensuit que M_t atteint son minimum pour $t=\mathbf{E}_\theta[\theta \mid X]$, de sorte que si la répartition *a priori* est connue, le problème d'estimation se résout jusqu'au bout.

Supposons maintenant que $h_0(\theta)$ est inconnue. Pour mesurer la qualité des estimations, on peut se servir de la quantité $\sup_\theta M_t(\theta)$, c'est-à-dire de la plus grande erreur possible.

En termes de $\sup_\theta M_{t^*}(\theta)$, une estimation t^* est meilleure qu'une estimation t si

$$\sup_\theta M_{t^*}(\theta) < \sup_\theta M_t(\theta).$$

Le *principe de minimax* régit le choix de l'estimation t^* pour laquelle $\sup_\theta M_{t^*}(\theta)$ est minimal. Si une telle estimation existe, alors

$$\sup_\theta M_{t^*}(\theta) \leqslant \sup_\theta M_t(\theta) \text{ pour tous les } t \neq t^*$$

et l'estimation t^* s'appelle *estimation du minimax*. Même si l'estimation du minimax peut être la moins bonne dans une grande partie de l'espace des paramètres, son avantage est que le supremum de son erreur est $\leqslant$ au supremum de l'erreur de toute autre estimation.

Exemple 8.11. L'estimation minimax de la probabilité inconnue p dans le schéma d'épreuves de Bernoulli de l'exemple 8.2 est ([51], [107])

$$\hat{p}^* = (r + \sqrt{n}/2)/(n + \sqrt{n}).$$

Les problèmes liés à l'approche bayesienne d'estimation sont examinés en détail dans les travaux [39], [141].

Conclusions

1. Un problème majeur de l'analyse statistique d'un système réel consiste à calculer (en fonction des données statistiques) les meilleures *valeurs approchées* (*les estimations statistiques*) d'un ou de plusieurs paramètres numériques caractérisant le fonctionnement de ce système. La possibilité d'obtenir des approximations « fonctionnelles » par le traitement statistique *d'une partie seulement* de la population générale analysée (c'est-à-dire d'un nombre restreint d'observations ou d'un échantillon) tient à la remarquable propriété de *stabilité statistique* des caractéristiques empiriques (cf. § 7.2).

2. L'estimation statistique se construit sous forme d'une fonction des observations et, *par nature*, *est une variable aléatoire*. Si l'on extrait un autre échantillon parent et que l'on porte les nouvelles valeurs empiriques dans la « fonction-estimation », on obtient en général une autre valeur approchée du paramètre estimé, autrement dit, les valeurs de l'estimation se dispersent de façon incontrôlable d'une expérience à l'autre (ici, d'un échantillon à l'autre).

3. La principale mesure de la précision d'une estimation statistique $\hat{\Theta}$ du paramètre inconnu Θ est l'écart-type de Θ par rapport à Θ,

c'est-à-dire $E(\hat{\Theta} - \Theta)^2$, et, dans le cas multidimensionnel, la matrice des covariances, des composantes de l'estimation vectorielle $\hat{\Theta}$. Il est évident que plus cette quantité (ou la dispersion généralisée de l'estimation $\hat{\Theta}$ dans le cas multidimensionnel) est petite, plus l'estimation est précise (*efficace*). Pour une vaste classe de populations générales, il existe une inégalité (l'inégalité de Rao-Cramer-Fréchet (8.12), (8.13)) qui définit le minimum $\Delta^2_{\min}$ (sur toutes les estimations possibles) de l'écart-type $E(\hat{\theta} - \theta)^2$ qu'il est impossible d'améliorer. Il est logique d'adopter ce minimum $\Delta^2_{\min}$ comme origine de la *mesure de l'efficacité* d'une estimation en définissant l'efficacité $e(\hat{\theta})$ de toute estimation $\hat{\theta}$ du paramètre par le rapport

$$e(\hat{\theta}) = \frac{\Delta^2_{\min}}{E(\hat{\theta} - \theta)^2}.$$

4. La *convergence* de l'estimation $\hat{\theta}$ (cf. § 8.1) assure sa stabilité statistique, c'est-à-dire sa convergence (stochastique) vers la véritable valeur du paramètre θ à mesure que la taille n de l'échantillon, sur le vu duquel cette estimation a été construite, croît. La propriété d'une estimation $\hat{\theta}$ d'être *sans biais* (cf. § 8.1) exprime que la moyennisation de toutes les valeurs possibles de cette estimation, obtenues sur le vu d'échantillons parents de même taille, nous conduit à la véritable valeur du paramètre estimé, soit $E\hat{\theta} = \theta$. Mais il n'est pas recommandé de rechercher toujours la réalisation de cette propriété : de peu de valeur pour les échantillons d'assez grandes tailles, elle risque d'appauvrir considérablement la classe des estimations dans laquelle on recherche la meilleure.

5. Vu que l'estimation $\hat{\theta}$ du paramètre inconnu θ est une variable aléatoire, il n'est pas sans intérêt de construire des *intervalles entiers* $\Delta\theta$ et, dans le cas multidimensionnel, des *régions entières*, qui avec une probabilité P proche de l'unité et donnée à l'avance recouvrent la véritable valeur du paramètre estimé θ, c'est-à-dire que $P\{\theta \in \Delta_P(\hat{\theta})\} = P$. Ces intervalles (resp. régions) s'appellent intervalles (resp. régions) *de confiance*. Il existe deux méthodes de construction de ces intervalles (resp. régions) : une méthode *exacte*, qui ne passe que pour une étroite classe de situations, et une méthode *approchée asymptotique*, plus répandue dans les applications statistiques ; cf. n° 8.6.5.

6. Les principales méthodes de construction des estimations statistiques sont :

la méthode du maximum de vraisemblance (cf. n° 8.6.1) ;

la méthode des moments (cf. n° 8.6.2) ;

la méthode des moindres carrés (cf. n° 8.6.3);

une méthode utilisant les observations « pondérées »: la censure, la troncature, les statistiques de rang (cf. n° 8.6.4).

Les diverses variantes de la méthode des observations « pondérées » font de plus en plus recette pour la stabilité des inférences statistiques par rapport aux éventuels écarts entre la loi réelle de la population générale étudiée et la loi modèle postulée.

7. La présence d'une information *a priori* sur le paramètre estimé θ, permettant d'associer à chaque valeur possible de θ une mesure probabiliste de sa certitude, c'est-à-dire des renseignements sur la loi de probabilité *a priori* du paramètre estimé, permet de préciser considérablement les estimations obtenues par les méthodes traditionnelles (méthode du maximum de vraisemblance, méthode des moments, etc.) en l'absence de cette information. Ces estimations sont construites par la *méthode bayesienne* (cf. n° 8.6.6) et sont dites bayesiennes.

CHAPITRE 9

TEST D'HYPOTHÈSES STATISTIQUES

La nécessité de formuler et de tester certaines hypothèses sur la nature ou la valeur des paramètres inconnus du schéma stochastique étudié se présente à divers stades de l'estimation statistique. Le chercheur avance, par exemple, l'hypothèse : « les observations analysées sont extraites d'un mélange de deux populations générales normales » ou encore « le vecteur des valeurs moyennes $A = (a^{(1)}, a^{(2)}, \ldots, a^{p})'$ des observations étudiées est égal au vecteur nul $0 = (0, 0, \ldots, 0)'$ », etc. On désignera les hypothèses par la lettre H. Notre objectif est de vérifier si l'hypothèse H ne contredit pas les données empiriques.

La comparaison de l'hypothèse émise aux données empiriques $X_1, X_2, \ldots, X_n$ est conduite à l'aide de tel ou tel test et s'appelle *test de l'hypothèse statistique.*

Le résultat de cette comparaison peut être soit *négatif* (les données empiriques contredisent l'hypothèse avancée et il faut alors y renoncer), soit *positif* (les données empiriques confirment l'hypothèse émise et celle-ci peut alors être prise pour éventuelle solution). A noter que si le résultat est positif, *cela ne veut encore pas dire que notre hypothèse est la meilleure et la seule possible : tout ce qu'on peut affirmer, c'est qu'elle ne contredit pas les données empiriques et il peut très bien exister d'autres hypothèses jouissant de cette propriété.* De sorte qu'une hypothèse H même statistiquement vérifiée n'est pas un fait absolu acquis une fois pour toute, mais seulement une proposition vraisemblable ne contredisant pas l'expérience.

Les hypothèses statistiques sont de plusieurs types. Passons-les en revue.

9.1. Principaux types d'hypothèses testées

9.1.1. Hypothèses relatives au type de loi de probabilité de la variable aléatoire étudiée. Quand on traite les observations

$$X_1, X_2, \ldots, X_n \tag{9.1}$$

de la variable aléatoire étudiée ξ, il est très important de saisir le mécanisme de formation des valeurs empiriques X_i, c'est-à-dire de sélectionner et de justifier une certaine fonction de répartition modèle $F_{\text{mod}}(X)$ (par exemple, parmi celles décrites au chapitre 6) qui

permette de décrire de façon adéquate la fonction de répartition $F_\xi(X)$. A ce stade de l'analyse, cela revient à éprouver des hypothèses du type

$$H: F_\xi(X) \equiv F_{\text{mod}}(X), \tag{9.2}$$

où la fonction modèle hypothétique peut être définie soit *de façon unique* (auquel cas $F_\xi(X) = F_0(X)$, où $F_0(X)$ est une fonction entièrement connue), soit *comme une fonction d'une famille paramétrique* (auquel cas $F_{\text{mod}}(X) = F(X; \Theta)$, où Θ est un paramètre de dimension k inconnu mais susceptible d'être estimé sur le vu de l'échantillon (9.1) à l'aide des méthodes développées au § 8.6).

La vérification des hypothèses (9.2) est réalisée à l'aide des *tests d'ajustement* et se base sur les diverses mesures de différenciation entre la fonction de répartition empirique analysée $\hat{F}_\xi^{(n)}(X)$ et la fonction modèle hypothétique $F_{\text{mod}}(X)$ (cf. § 11.1).

9.1.2. Hypothèses d'homogénéité de deux ou plusieurs échantillons ou de certaines caractéristiques des populations analysées. Les plus typiques des problèmes de cette nature sont caractérisés par la situation suivante. Supposons qu'on dispose de l « tranches » de données empiriques de type (9.1):

$$\begin{array}{l} X_{11}, X_{12}, \ldots, X_{1n_1}; \\ X_{21}, X_{22}, \ldots, X_{2n_2}; \\ \ldots\ldots\ldots\ldots\ldots \\ X_{l1}, X_{l2}, \ldots, X_{ln_l} \end{array} \tag{9.3}$$

Ces tranches auraient pu se former de façon naturelle au cours d'un sondage (par exemple, elles auraient pu être enregistrées à des dates et en des lieux différents). En désignant par $F_j(X)$ la fonction de répartition des observations du j-ième échantillon et en attribuant le même indice à toutes les caractéristiques empiriques et théoriques de cette loi (les moyennes $\hat{a}_j(n_j)$ et a_j; les variances $\hat{\sigma}_j^2(n_j)$ et σ_j^2, etc), on peut mettre les principales hypothèses d'homogénéité sous la forme:

$$H_F: F_1(X) \equiv F_2(X) \equiv \ldots \equiv F_l(X); \tag{9.3a}$$

$$H_a: a_1 = a_2 = \ldots = a_l; \tag{9.3b}$$

$$H_\sigma: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2. \tag{9.3c}$$

Si le résultat du test de ces hypothèses est positif, on dit que les caractéristiques empiriques correspondantes (par exemple, $\hat{a}_1(n_1)$, $\hat{a}_2(n_2)$, ..., $\hat{a}_l(n_l)$) sont *non significativement* différentes.

Signalons le cas particulier des hypothèses (9.3a) où le nombre des échantillons $l = 2$ et l'un des échantillons est composé d'un petit nombre d'observations (une, pour fixer les idées). Sous cette forme,

tester des hypothèses (9.3a) revient à *tester l'anomalie d'une ou de plusieurs observations*. Les tests des hypothèses de type (9.3a), (9.3b) et (9.3c) sont décrits aux §§ 11.2 et 11.5.

9.1.3. Hypothèses concernant les valeurs numériques des paramètres de la population générale étudiée. Supposons, par exemple, que les observations (9.1) sont des valeurs d'un paramètre d'un article, obtenues par des mesures effectuées sur n articles tirés au hasard dans la production en série d'une machine et soit a_0 la valeur nominale de ce paramètre. Chaque valeur X_i peut naturellement différer de la valeur nominale a_0. Il est évident que pour vérifier que cette machine est correctement réglée, il faut s'assurer que la valeur moyenne du paramètre des unités produites par cette machine correspond à la valeur nominale, c'est-à-dire éprouver une hypothèse du type

$$H: \mathbf{E}\xi = a_0. \tag{9.4}$$

Nous sommes conduits à une hypothèse de cette nature si l'on tente de vérifier que le coefficient de corrélation empirique $\hat{r}\,(x^{(1)}, x^{(2)})$, construit sur l'ensemble des observations à deux dimensions $X_i = (x_i^{(1)}, x_i^{(2)})$, $i = 1, 2, \ldots, n$ (cf. n° 5.6.7), diffère statistiquement de façon non significative du zéro. Compte tenu de la caractéristique théorique envisagée, cette hypothèse peut être mise sous la forme

$$H: r\,(x^{(1)}, x^{(2)}) = 0. \tag{9.4'}$$

Dans le cas général, les hypothèses de cette nature sont de la forme:

$$H_0: \Theta = \Delta_0, \tag{9.4''}$$

où Θ est un paramètre dont dépend la loi étudiée, Δ_0 le domaine de ses valeurs hypothétiques concrètes, domaine qui peut très bien n'être composé que d'un point.

Le n° 11.2.4 est consacré à l'un des cas particuliers des hypothèses (9.4″): au test de l'hypothèse $\beta_1 = 0$ (c'est-à-dire au *test de la symétrie* de la loi). Les *hypothèses d'indépendance et de stationnarité des observations* (cf. plus bas n° 9.1.5, ainsi que le § 11.3) se ramènent au type (9.4″).

9.1.4. Hypothèses concernant le type de dépendance entre les composantes de la variable multidimensionnelle étudiée. Autant dans l'étude de la loi de probabilité des observations traitées il est important de choisir correctement le modèle de loi correspondant (cf. n° 9.11), autant dans l'étude de la dépendance statistique, par exemple de la composante $x^{(2)}$ par rapport à la composante $x^{(1)}$, soit $X = (x^{(1)}, x^{(2)})'$, il est important de tester l'*hypothèse concernant la forme générale de cette dépendance*. Par exemple, l'hypothèse que

$x^{(2)}$ et $x^{(1)}$ sont reliées par une relation de régression *linéaire* (cf. n° 8.6.3, ainsi que le § 3.5), soit

$$H: \mathrm{E}\,(x^{(2)} \mid x^{(1)} = x) = b_0 + b_1 x, \tag{9.5}$$

où b_0 et b_1 sont des paramètres inconnus du modèle.

Les tests statistiques utilisés pour éprouver les hypothèses de type (9.5) sont souvent appelés *tests d'adéquation* (cf. par exemple [7]).

9.1.5. Hypothèses concernant l'indépendance et la stationnarité des observations traitées. La question de savoir si l'on peut traiter des observations (9.1) comme indépendantes ou à la rigueur stationnaires, c'est-à-dire suivant la même loi tout au long du sondage, n'est pas dénuée de sens dans la mesure en effet où elle conditionne le choix de la meilleure méthode de traitement de ces observations. Les hypothèses correspondantes peuvent être formalisées en fonction des objectifs de l'analyse, par exemple comme suit :

$$H: \mathrm{E}x_i = a = \text{const}, \quad i = 1, 2, \ldots, n; \tag{9.6}$$

$$H: r\,(x_i, x_{i+1}) = 0, \; i = 1, 2, \ldots, n-1, \tag{9.6'}$$

c'est-à-dire qu'à strictement parler les hypothèses de ce type peuvent être traitées comme un cas particulier d'hypothèses (9.4″). Les tests correspondants sont décrits dans le § 11.3.

9.2. Principe logique général d'un test statistique

Les tests statistiques diffèrent considérablement par leur finalité et le caractère des problèmes résolus. Mais ils sont tous construits suivant le même principe logique que l'on peut décrire de la manière suivante.

1. On avance une hypothèse H_0.

2. On définit le seuil de signification α du test. En effet, toute décision statistique, c'est-à-dire toute décision prise sur la base d'un nombre *fini* d'observations est adoptée avec un certain risque d'erreur dans un sens comme dans l'autre. Dans un nombre α peu élevé de cas, l'hypothèse H_0 peut être rejetée alors qu'elle est vraie (on parle d'un risque (ou d'une erreur) de première espèce) et, inversement, dans un nombre β peu élevé de cas, elle peut être adoptée alors qu'elle est fausse, la vraie étant une hypothèse concurrente H_1 (on parle d'un risque (ou d'une erreur) de deuxième espèce). En fixant la taille de l'échantillon, on peut choisir à notre convenance la probabilité de l'un ou l'autre de ces risques. Si l'on peut accroître autant que l'on veut le volume de l'échantillon, on peut rendre arbitrairement petites les probabilités des risques α et β pour toute hypothèse concurrente H_1. En particulier, si le volume de l'échantillon est fixe, on se donne généralement la probabilité α de rejet à tort de l'hypothèse H_0, appelée souvent hypothèse *de base* ou *nulle*. Cette proba-

bilité de rejet à tort de l'hypothèse de base s'appelle *seuil de signification* ou encore *niveau du test*. Le choix de la valeur du seuil de signification α dépend de la comparaison des pertes subies en cas de décision erronée dans un sens ou dans l'autre : plus les pertes entraînées par le rejet à tort de H_0 sont lourdes et plus α doit être pris petit. Mais comme cette comparaison est dans la plupart des cas très difficile à réaliser (souvent, il est difficile de dire dans quel sens l'erreur est la plus dangereuse), on fait appel en général à certaines valeurs standards du seuil de signification. Ces valeurs sont : $\alpha = 0{,}1$; 0,05 ; 0,025 ; 0,01 ; 0,005 ; 0,001. La plus répandue de ces valeurs est la valeur $\alpha = 0{,}05$. Elle exprime qu'en moyenne dans 5 % des cas nous rejetterons à tort l'hypothèse avancée.

3. Soit donnée une fonction des observations $\gamma^{(n)} = \gamma(X_1, \ldots, X_n)$ (appelée *statistique critique*). Comme toute fonction des observations, la statistique critique $\gamma^{(n)}$ est une variable aléatoire (cf. § 8.1) qui, si H_0 est vraie, suit une loi de probabilité bien étudiée (tabulée) de densité $f_{\gamma^{(n)}}(u)$.

Le principe de construction de la statistique critique (dit *principe du rapport de vraisemblance*) est décrit au paragraphe suivant. Etudions dans les grandes lignes la signification de cette statistique : en principe, elle mesure le désaccord entre les données empiriques (9.1) et l'hypothèse testée H_0. Ainsi, dans les hypothèses envisagées au n° 9.1.1, la statistique critique $\gamma^{(n)}$ mesure la différence entre la fonction de répartition empirique $\hat{F}^{(n)}(X)$ et la fonction modèle hypothétique $F_{\text{mod}}(X)$. Dans les hypothèses du n° 9.1.2, la quantité $\gamma^{(n)}$ mesure l'écart entre les caractéristiques empiriques dans les divers échantillons ; dans les hypothèses du n° 9.1.3. enfin, elle rend compte de l'écart entre les caractéristiques empiriques et leurs valeurs théoriques, etc.

4. Dans les tables de la répartition de densité $f_{\gamma^{(n)}}(u)$, on repère le point $\gamma_{\alpha/2}^{(\min)}$ de pourcentage $100\left(1-\frac{\alpha}{2}\right)$ et le point $\gamma_{\alpha/2}^{(\max)}$ de pourcentage $100\alpha/2$ (cf. § 5.6). Ces points subdivisent le domaine des valeurs possibles de la variable aléatoire $\gamma^{(n)}$ en trois régions (cf. fig. 9.1) : la *région (I) des valeurs invraisemblablement petites*, la *région (III) des valeurs invraisemblablement grandes* et, enfin, la *région (II) des valeurs vraisemblables* ou *naturelles* (*sachant que l'hypothèse H_0 est vraie*). Si l'on craint que les erreurs qui entachent l'hypothèse avancée soient unilatérales, c'est-à-dire que les valeurs de la statistique critique sont seulement « trop grandes » ou seulement « trop petites », on ne cherchera qu'un seul point de pourcentage : soit le point $\gamma_{\alpha}^{(\min)}$ de pourcentage $100(1-\alpha)$ qui partagera le domaine des valeurs de $\gamma^{(n)}$ en deux régions : une région de valeurs invraisemblablement petites et une région de valeurs vraisemblables ; soit le point $\gamma_{\alpha}^{(\max)}$ de pourcentage 100α qui divisera le domai-

ne des valeurs de $\gamma^{(n)}$ en deux régions : une région des valeurs invraisemblablement grandes et une région des valeurs vraisemblables.

5. Enfin, on porte les données empiriques $X_1, \ldots, X_n$ dans la fonction $\gamma^{(n)}$ et on calcule la *valeur prise par* $\gamma^{(n)}$. Si cette valeur appartient à la région des valeurs vraisemblables, l'hypothèse H_0

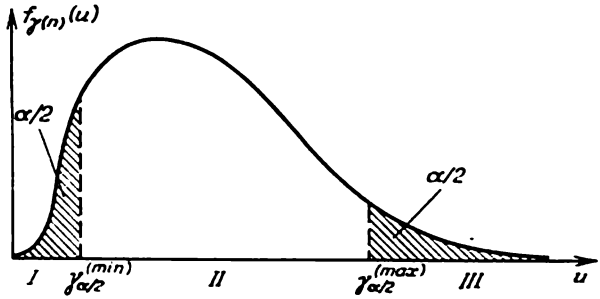


Fig. 9.1. Représentation graphique de la densité de la statistique critique γ_n et mise en évidence des régions d'acceptation (*II*) et de rejet (*I* et *III*), sachant que H_0 est vraie

ne contredit pas les données empiriques. Dans le cas contraire, c'est-à-dire si $\gamma^{(n)}$ est trop petite ou trop grande, on conclut que $\gamma^{(n)}$ ne suit pas la loi de densité $f_{\gamma^{(n)}}(u)$ (il est aisé de comprendre que cette conclusion est établie avec une probabilité d'erreur égale à α), ce qui exprime que notre hypothèse H_0 est fausse et il faut donc y renoncer.

Donc, la décision suggérée par un test statistique peut être fausse aussi bien dans le cas où l'hypothèse H_0 est rejetée (avec une probabilité α) que dans le cas où elle est acceptée (avec une probabilité β). Les probabilités α et β des décisions fausses s'appellent respectivement *risque* (*ou erreur*) *de première* et *de deuxième espèce*, et la quantité $1 - \beta$, *puissance du test*. Il est évident que de deux tests caractérisés par un même risque de première espèce α, on préférera celui dont le risque de deuxième espèce β est plus petit (ou dont la puissance est la plus grande).

Si l'hypothèse H_0 consiste à conjecturer que la valeur d'un paramètre Θ est *très exactement* égale à une valeur donnée Θ_0 (cf. plus haut les hypothèses envisagées au n° 9.1.3), on dit alors qu'elle est *simple*. Dans tous les autres cas, elle est *multiple*.

9.3. Construction d'un test statistique ; principe du rapport de vraisemblance

Voyons comment obtenir ces fonctions d'observations (les statistiques critiques $\gamma^{(n)}$) sur le vu des valeurs desquelles on décide si l'hypothèse testée correspond ou non aux valeurs empiriques (9.1).

9.3.1. Essence du principe du rapport de vraisemblance. Pour comprendre le principe général qui conduit à la construction des meilleurs (des plus puissants pour un seuil de signification donné) tests, revenons à l'exemple du salaire (cf. n° 8.6.1) et à la figure 8.2. Dans cet exemple, on a étudié la loi lognormale du salaire ξ de personnes actives prélevées dans une certaine population avec les données initiales suivantes (les salaires de trois personnes sondées): $x_1 = 190$ roubles; $x_2 = 175$ roubles et $x_3 = 205$ roubles. Soit à tester l'hypothèse (simple) concernant la moyenne de la variable aléatoire normale $\ln \xi$:

$$H_0\colon \mathrm{E} \ln \xi = 5{,}240 = a_0,$$

l'hypothèse concurrente étant

$$H_1\colon \mathrm{E} \ln \xi = 5{,}443 = a_1.$$

On voit sur la figure 8.2 que l'hypothèse H_0 ne contredit pas les données initiales (bien plus, ces données sont plus vraisemblables pour l'hypothèse H_0, alors qu'elles sont peu vraisemblables si H_1 est vraie).

Dans le cas général, on se fait une idée de la vraisemblance des observations $X_1, \ldots, X_n$ (relativement à l'hypothèse testée et à l'hypothèse concurrente) en comparant les fonctions de vraisemblance respectives (cf. formule (8.5)), plus exactement, en composant leur rapport

$$\gamma^{(n)} = \frac{L_{H_1}(X_1, \ldots, X_n; \Theta)}{L_{H_0}(X_1, \ldots, X_n; \Theta)} = \frac{L(X_1, \ldots, X_n; \Theta_1)}{L(X_1, \ldots, X_n; \Theta_0)}, \tag{9.7}$$

où L_{H_1} et L_{H_0} sont les valeurs des fonctions de vraisemblance des observations $X_1, \ldots, X_n$ calculées sous la condition que soient vraies respectivement les hypothèses H_1: $\Theta = \Theta_1$ et H_2: $\Theta = \Theta_0$. Il est évident que plus les observations sont vraisemblables dans le cas de l'hypothèse H_0, plus la fonction de vraisemblance L_{H_0} prend des valeurs élevées et plus la quantité $\gamma^{(n)}$ est petite. Si $f_{\gamma^{(n)}}(u)$ est la densité de probabilité de la statistique $\gamma^{(n)}$ sachant que l'hypothèse H_0 est vraie, la construction d'un test de l'hypothèse H_0 de niveau α donné revient à déterminer le point γ_α de pourcentage 100α de la loi de densité $f_{\gamma^{(n)}}(u)$ et à mettre en œuvre la règle suivante:

si $\gamma^{(n)} > \gamma_\alpha$ l'hypothèse H_0 est rejetée avec un risque α, puisque, conformément à la loi de densité $f_{\gamma^{(n)}}(u)$ et sous réserve que l'hypothèse H_0 soit vraie, l'événement $\{\gamma^{(n)} > \gamma_\alpha\}$ peut se réaliser avec la probabilité α, c'est-à-dire que

$$\int_{\gamma_\alpha}^{\infty} f_{\gamma^{(n)}}(u)\, du = \alpha\,;$$

si $\gamma^{(n)} < \gamma_\alpha$, l'hypothèse H_0 est acceptée. (9.8)

Les tests basés sur les statistiques $\gamma^{(n)}$ de la forme (9.7) et sur la procédure (9.8) s'appellent *tests du rapport de vraisemblance*. Ces tests sont préférés aux autres pour leur réalisabilité et les propriétés suivantes (qui sont valables dans une classe assez vaste de situations, cf. par exemple [40]).

1. Les tests du rapport de vraisemblance sont les plus puissants (lemme de Neyman-Pearson).

2. La densité $f_{\gamma^{(n)}}(u)$ de la loi de probabilité de la statistique critique $\gamma^{(n)}$ se calcule généralement sans peine à partir de la fonction de vraisemblance L de la variable aléatoire analysée.

La généralisation du problème de test de l'hypothèse concernant la moyenne d'une variable aléatoire normale ξ (sachant que la variance σ^2 est connue) nous donne

$$L_{H_j} = L(x_1, \ldots x_n\,;\ \Theta_j) = L(x_1, \ldots, x_n\,;\ a_j, \sigma^2) =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a_j)^2} \quad (j = 0,\ 1),$$

de sorte que

$$\gamma^{(n)} = \frac{L_{H_1}}{L_{H_0}} = e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i-a_1)^2-(x_i-a_0)^2]} = e^{c\frac{(\bar{x}-a_0)\sqrt{n}}{\sigma} - \frac{1}{2}c^2},$$

où $c = \sqrt{n}\,(a_1 - a_0)\cdot\frac{1}{\sigma}$. Supposons pour fixer les idées que $a_1 > a_0$ (dans l'exemple ci-dessus $n = 3$; $\sigma^2 = 0{,}16$; $a_1 = 5{,}443$ et $a_0 = 5{,}240$). Alors $c > 0$, et si l'on pose

$$\gamma_\alpha = e^{cu_{Q(\alpha)} - \frac{1}{2}c^2},$$

où $Q(\alpha) = 100\cdot 2\alpha$ % et u_Q est le point de pourcentage $100Q$ de la loi normale réduite, l'inégalité

$$e^{c\frac{(\bar{x}-a_0)\sqrt{n}}{\sigma} - \frac{1}{2}c^2} > \gamma_\alpha$$

sera réalisée sur l'ensemble des échantillons $(x_1, \ldots, x_n)$ tels que $\sqrt{n}\,(\bar{x}\,(n) - a_0)\,\frac{1}{\sigma} > u_{Q(\alpha)}$, ou, ce qui revient au même,

$$\bar{x}\,(n) \geqslant a_0 + u_{Q(\alpha)} \cdot \frac{\sigma}{\sqrt{n}}.$$

Cette règle de test de l'hypothèse ne dépend pas de la valeur concurrente du paramètre a_1, donc est la plus puissante (eu égard au lemme de Neyman-Pearson) *pour toutes les valeurs* concurrentes *possibles* du paramètre $a_1 > a_0$ on dit encore qu'elle est *uniformément la plus puissante.*

9.3.2. Test d'une hypothèse simple par le logarithme du rapport de vraisemblance. Supposons qu'on sache que des observations $X_1, \ldots, X_n$ peuvent être traitées comme un échantillon indépendant prélevé dans une population générale dont la fonction de répartition appartient à une famille de fonctions de répartition $F\,(X;\,\Theta)$, où Θ est un paramètre à k dimensions. On demande de tester l'hypothèse $\Theta = \Theta_0$ (hypothèse (9.4″) du n° 9.1.3). Considérons le test

$$\tilde{\gamma}^{(n)} = -2 \ln \{L\,(X_1, \ldots, X_n;\,\Theta_0)/L\,(X_1, \ldots, X_n;\,\hat{\Theta})\}, \quad (9.9)$$

où Θ est une estimation (introduite au n° 8.6.1) du maximum de vraisemblance du paramètre Θ sur le vu des observations $X_1, \ldots, X_n$. Si l'on impose à la famille $F\,(X;\,\Theta)$ et à la valeur Θ_0 des conditions supplémentaires garantissant la réalisation des bonnes propriétés des estimations du maximum de vraisemblance (cf. n° 8.6.1), la variable $\tilde{\gamma}^{(n)}$ suivra asymptotiquement (pour $n \to \infty$) une loi du χ^2 à k degrés de liberté (cf. n° 6.2.1).

A titre d'exemple d'application du test $\tilde{\gamma}^{(n)}$, revenons au problème du test de l'hypothèse concernant la moyenne d'une population normale, envisagé au n° 9.3.1. Dans ces notations, en tenant compte du fait que l'estimation du maximum de vraisemblance du paramètre a_0 est $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, on obtient

$$\tilde{\gamma}^{(n)} = 2\left(\sum_{i=1}^{n} \frac{(x_i - a_0)^2}{2\sigma^2} - \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{2\sigma^2}\right) = n\,(\bar{x} - a_0)^2/\sigma^2.$$

Comme $\bar{x}$ est une variable aléatoire normale de moyenne a_0 et de variance $\frac{\sigma^2}{n}$, il s'ensuit que $\tilde{\gamma}^{(n)}$ est une variable du χ^2 à un degré de liberté.

A titre de deuxième exemple, considérons le test de l'hypothèse H_0: $p_i = p_{0i}$, $i = 1, 2, \ldots, l$, concernant les valeurs des para-

mètres $p_1, p_2, \ldots, p_l$ de la loi polynomiale du n° 6.1.4. Vu que les estimations du maximum de vraisemblance du paramètre p_i sont les rapports $\nu_i(n)/n$ (cf. n° 8.6.1), il s'ensuit que la variable

$$\tilde{\gamma}^{(n)} = 2 \sum_{i=1}^{l} \nu_i(n) (\ln \nu_i(n) - \ln np_{0i})$$

suit une loi asymptotique du χ^2 à $l - 1$ degrés de liberté.

9.3.3. Test d'une hypothèse multiple. Considérons une modification du test (9.9) pour le cas où dans l'hypothèse sont concrétisés une partie seulement des paramètres. Supposons que $\Theta = (\theta_1, \ldots, \theta_l)$ est le vecteur des paramètres inconnus de la loi et que l'hypothèse est

$$H_{01}: \theta_j = \theta_{0j}, \quad j = 1, \ldots, r < l. \tag{9.10}$$

Il est commode de diviser le vecteur Θ en deux parties: $\Theta_1 = (\theta_1, \ldots, \theta_r)$ et $\Theta_2 = (\theta_{r+1}, \ldots, \theta_l)$. Désignons par $\tilde{\Theta}_2$ l'estimation du maximum de vraisemblance de Θ_2 sur le vu de $X_1, \ldots, X_n$ sachant que $\Theta_1 = \Theta_{01}$ et par $(\hat{\Theta}_1, \hat{\Theta}_2)$ l'estimation du maximum de vraisemblance de (Θ_1, Θ_2). Le test de l'hypothèse H_{01} est défini par la statistique

$$\hat{\gamma}^{(n)} = -2 \ln \{L(X_1, \ldots, X_n; \Theta_{01}, \tilde{\Theta}_2)/L(X_1, \ldots, X_n; \hat{\Theta}_1, \hat{\Theta}_2)\}. \tag{9.11}$$

On démontre que la quantité $\hat{\gamma}^{(n)}$ suit asymptotiquement (lorsque $n \to \infty$) une loi du χ^2 à r degrés de liberté si sont remplies des conditions supplémentaires assurant la réalisation des bonnes propriétés des estimations du maximum de vraisemblance (cf. n° 8.6.1).

9.4. Critères de « qualité » d'un test statistique

Les critères définissant la précision d'un test statistique d'hypothèses simples ou multiples de type (9.4″) sont:

$\alpha(\Theta)$ la probabilité de rejeter l'hypothèse de base H_0 sous réserve que la véritable valeur du paramètre « testé » soit Θ; la quantité $1 - \alpha(\Theta)$ s'appelle *caractéristique opérationnelle* du test, quant à la valeur $\alpha(\Theta_0)$ dans le test de l'hypothèse simple H_0: $\Theta = \Theta_0$, ce n'est autre que le niveau de ce test (le risque de première espèce);

$\beta(\Theta)$ la probabilité de rejeter l'hypothèse concurrente (de l'hypothèse de base) sous réserve que la véritable valeur du paramètre « testé » soit égale à Θ; la quantité $1 - \beta(\Theta)$ s'appelle puissance (ou fonction de puissance) du test, quant à $\beta(\Theta_1)$, ce n'est autre que le risque de deuxième espèce dans le problème du test de l'hypothèse simple H_0: $\Theta = \Theta_0$ contre l'hypothèse concurrente H_1: $\Theta = \Theta_1$.

Discutons ces caractéristiques et les propriétés du test qui en dépendent.

Soient Γ_n le domaine des valeurs possibles de la statistique critique $\gamma^{(n)}$, $\Gamma_n^{H_0}$ et $\Gamma_n^{H_1}$ les régions, décrites au § 9.2, respectivement des valeurs « vraisemblables » et « invraisemblables » de $\gamma^{(n)}$ (sachant que l'hypothèse H_0 est vraie). Il est alors évident que

$$\begin{cases} \alpha_n(\Theta) = P_\Theta\{\gamma^{(n)} \in \Gamma_n^{H_1}\} = \int\limits_{\Gamma_n^{H_1}} f_{\gamma^{(n)}}(u, \Theta)\, du; \\ \beta_n(\Theta) = 1 - \alpha_n(\Theta) = P_\Theta\{\gamma^{(n)} \in \Gamma_n^{H_0}\} = \int\limits_{\Gamma_n^{H_0}} f_{\gamma^{(n)}}(u; \Theta)\, du, \end{cases} \tag{9.12}$$

où $P_\Theta\{A\}$ et $f_{\gamma^{(n)}}(u; \Theta)$ sont respectivement la probabilité de l'événement A et la densité de probabilité de la statistique critique $\gamma^{(n)}$, calculée sous l hypothèse que la véritable valeur du paramètre testé soit égale à Θ.

Dans le cas d'hypothèses paramétriques de la forme (9.4″) au seuil de signification α_0 donné, le test $\{\gamma^{(n)}, \Gamma_n^{H_0}\}$ est dit sans *biais* si

$$\begin{cases} \alpha_n(\Theta) \leqslant \alpha_0 & \text{pour tous les } \Theta \in \Delta_0; \\ \alpha_n(\Theta) > \alpha_0 & \text{pour tous les } \Theta \notin \Delta_0. \end{cases}$$

Enfin, le test $\{\gamma^{(n)}, \Gamma_n^{H_0}\}$ est dit *convergent* si $\lim\limits_{n\to\infty} \alpha_n(\Theta) = 1$ pour tous les $\Theta \notin \Delta_0$.

La dernière relation exprime, en particulier, que la puissance $1 - \beta_n(\Theta)$ du test convergent tend vers l'unité (lorsque $n \to \infty$) pour toute valeur Θ n'appartenant pas au domaine Δ_0 des valeurs hypothétiques (correspondant à l'hypothèse H_0) du paramètre.

De (9.12) il s'ensuit visiblement que pour toute taille n fixe de l'échantillon, la modification du test dans le sens d'une réduction du niveau α (c'est-à-dire la restriction du domaine $\Gamma_n^{H_1}$) est liée à un accroissement simultané du risque de deuxième espèce et, dans le cas général, à une réduction des valeurs de la puissance $1 - \beta(\Theta)$ (puisque dans ce cas la région $\Gamma_n^{H_0}$ de rejet de l'hypothèse concurrente H_1 s'élargit). Réciproquement : la modification d'un test (dans toute classe fixe de tests, y compris dans la classe des tests les plus puissants) dans le sens de l'accroissement de sa puissance est liée (pour n fixe) à une inévitable augmentation simultanée de son niveau.

Dans le même temps, en faisant croître indéfiniment la taille de l'échantillon (c'est-à-dire en faisant tendre $n \to \infty$), on peut faire

prendre aux risques

$$\alpha_n^* = \begin{cases} \alpha_n(\Theta^*), & \text{où } \Theta^* \in \Delta_0^* \subset \Delta_0, \text{ ou} \\ \max\limits_{\Theta \in \Delta_0} \alpha_n(\Theta); \end{cases}$$

$$\beta_n^* = \begin{cases} \beta_n(\Theta^{**}), & \text{où } \Theta^{**} \notin \Delta_0, \text{ ou} \\ \max\limits_{\Theta \notin \Delta_0} \beta_n(\Theta) \end{cases}$$

des valeurs aussi petites que l'on veut.

Pour les échantillons de grande taille (c'est-à-dire asymptotiquement par rapport à $n \to \infty$), il existe des relations liant entre elles les caractéristiques α_n^*, β_n^* et n (cf. par exemple [77]). Appesantissons-nous sur une relation utile qui nous permettra, en particulier, de déterminer la taille $n(\alpha, \beta; \rho)$ d'un échantillon dans un test du maximum de vraisemblance (de Neyman-Pearson) entre deux hypothèses simples:

$$\begin{aligned} &H_0: \text{ l'échantillon est extrait d'une population générale de densité } f(X; \Theta_0); \\ &H_1: \text{ l'échantillon est extrait d'une population générale de densité } f(X; \Theta_1) \end{aligned} \tag{9.13}$$

avec des risques de première et de deuxième espèces, inférieurs respectivement à des valeurs données α et β (la quantité $\rho = \rho(H_0, H_1)$ caractérise la « distance » entre les hypothèses H_0 et H_1 et est définie par la formule $\rho(H_0, H_1) = \int \ln \frac{f(X; \Theta_1)}{f(X; \Theta_0)} [f(X; \Theta_1) - f(X; \Theta_0)]\, dX$, où l'intégration est étendue au domaine tout entier des valeurs possibles de la variable aléatoire X, $f(X; \Theta)$ sa densité de probabilité). Dans [5], on prouve que dans une classe assez vaste de cas, quand on procède à des tests entre des hypothèses simples voisines (c'est-à-dire pour de petites valeurs de ρ), on peut se servir de la formule (asymptotique) approchée

$$n(\alpha; \beta; \rho) \approx \frac{(u_{1-\alpha} + u_{1-\beta})^2}{\rho(H_0, H_1)}, \tag{9.14}$$

dans laquelle u_q est comme toujours le quantile d'ordre q de la loi normale réduite (cf., par exemple, tableau 1.3 dans [27]).

Remarque. Attirons l'attention du lecteur sur l'apparition inévitable de deux effets « néfastes » à la théorie des tests d'hypothèses statistiques: les effets « *trop petite taille* » et « *trop grande taille* » de l'échantillon.

L'effet « trop petite taille » de l'échantillon consiste en ce qui suit : si le niveau α du test est donné et si le nombre n d'observations sur le vu desquelles est prise la décision est petit, la puissance du test, c'est-à-dire la probabilité de rejeter l'hypothèse de base H_0

à juste titre, est très petite (la formule (9.14) nous donne une idée approximative de l'interdépendance des quantités α, β et n). Il existe deux remèdes à cette situation : soit augmenter la taille n de l'échantillon, soit augmenter un peu le niveau α, ce qui entraînera une diminution de β (c'est-à-dire un accroissement de la puissance $1 - \beta$).

Pour expliquer l'effet « trop grande taille » de l'échantillon, nous ferons appel à une réflexion de Berkson (Journ. Amer. Statist. Assoc., 33 (1938), p. 526) : « Personne en réalité n'estime qu'une hypothèse se réalise exactement : nous ne faisons que construire un modèle abstrait d'événements réels qui dans une certaine mesure s'écarte nécessairement de la vérité. Mais nous constatons qu'un échantillon de très grande taille infirme dans ce cas presque sûrement (c'est-à-dire avec une probabilité qui tend vers l'unité lorsque n croît indéfiniment) *) notre hypothèse pour tout niveau α donné. »

On dirait un cercle vicieux : pour les échantillons de faible taille, les conclusions ne sont pas fiables statistiquement, pour les échantillons de trop grande taille, le résultat est connu sans ambiguïté à l'avance. Les auteurs de cet ouvrage ont plus d'une fois été les témoins amusés de la perplexité, dans laquelle l'effet des grands n jetait les chercheurs qui tentaient au moyen d'un test de trouver une loi modèle $F_{\text{mod}}(x)$ convenable pour décrire la population générale étudiée et qui aboutissaient invariablement au résultat contraire (c'est-à-dire au rejet de l'hypothèse avancée).

Pour éviter l'effet des grands n, il faut relier les caractéristiques de précision du test (le niveau α et le risque de deuxième espèce β) à la taille n de l'échantillon : le gain en « sensibilité » du test, réalisé grâce à l'accroissement de n, doit être utilisé pour réduire aussi bien α que β. En particulier, si α diminue lorsque n croît, les petits écarts par rapport à H_0 ne conduiront déjà plus au rejet obligatoire de cette hypothèse : la probabilité de ce rejet dépendra de la « vitesse » de décroissance de α (lorsque n croît).

9.5. Procédure séquentielle de prise de décision (tests séquentiels)

9.5.1. Procédure séquentielle d'observation. Si le nombre d'observations, sur le vu desquelles le statisticien est appelé à prendre une décision, n'est pas fixé à l'avance mais est tributaire des résultats enregistrés à chaque étape de l'expérience, on dit qu'on utilise une *procédure séquentielle d'observation*. Vu que les résultats des observations à chaque étape de l'expérience constituent un échantillon aléatoire de la population générale, donc, sont aléatoires, l'instant

*) Ceci résulte de la propriété du test d'être sans biais (cf. plus haut).

d'arrêt des observations (qui dépend de ces résultats) est aussi une variable aléatoire.

L'idée d'utiliser la procédure séquentielle d'observation est apparue pour la première fois au cours de l'établissement des plans économiques de contrôle statistique par échantillonnage de la qualité d'une production *). Il s'agissait de vérifier sur un échantillon que le pourcentage p de pièces défectueuses produites en série ne dépassait pas un certain niveau p_0 (le niveau limite tolérable). Les auteurs du travail cité ont proposé à cet effet une procédure de contrôle par un double échantillonnage. Dans une première étape, ils ont prélevé un échantillon de taille n_1 dont ils ont déterminé le pourcentage $\hat{p}_1(n_1)$ de pièces défectueuses. La décision de contrôler un deuxième échantillon se prend sur le vu du résultat obtenu : en gros, si le pourcentage $\hat{p}_1(n_1)$ de pièces défectueuses dans le premier échantillon est bien plus petit (ou bien plus grand) que le niveau limite p_0, il n'est point besoin de considérer un deuxième échantillon et l'on accepte l'hypothèse $p \leqslant p_0$ (resp. l'hypothèse concurrente $p > p_0$) ; si le pourcentage $\hat{p}_1(n_1)$ diffère peu du niveau p_0 (ou comme on dit encore se trouve dans la « région d'indifférence » ou « région d'indétermination »), on prend la décision de poursuivre les observations et, en particulier, de prélever un deuxième échantillon de taille n_2.

Cette méthode fournit un gain appréciable (en moyenne) sur le nombre d'observations nécessaires pour tester les hypothèses à un niveau α et une puissance $1 - \beta$ donnés par rapport au test de Neyman-Pearson qui est le meilleur (le plus puissant, cf. n° 9.3.1) de tous les tests basés sur la procédure classique d'observation (c'est-à-dire sur des échantillons *de taille n donnée a priori*). Donc, la procédure séquentielle d'observation s'impose lorsque chaque observation est d'un coût élevé ou d'accès difficile et à condition que le chercheur soit en état de mettre cette procédure en œuvre (ce qui n'est pas toujours le cas).

Aux caractéristiques $\alpha(\Theta)$ et $\beta(\Theta)$ de la qualité d'un test (cf. § 9.4) il faut ajouter dans le cas d'un test séquentiel la *taille moyenne de l'échantillon* $E_\Theta \nu(\alpha, \beta)$ nécessaire pour vérifier les hypothèses de la forme (9.4″) avec des caractéristiques (α, β) données.

9.5.2. Test séquentiel du rapport de vraisemblance (test de Wald) et ses propriétés. La construction d'un test statistique dans le cas où la taille n de l'échantillon est fixée (cf. n° 9.3.1) revient à partager le domaine des valeurs possibles de la statistique critique $\gamma^{(n)} = \gamma(X_1, \ldots, X_n)$ en deux régions : la région des valeurs vraisemblables et la région des valeurs invraisemblables (sachant que l'hypo-

*) Dodge H.F., Romig H.G. *A method of sampling inspection.* — The Bell System Tech. Journ., 8 (1929), 613-631.

thèse de base H_0 est vraie). Si une valeur donnée de $\gamma(X_1, \ldots, X_n)$ tombe dans la région des valeurs invraisemblables, l'hypothèse de base est rejetée.

Les tests séquentiels, c'est-à-dire les tests basés sur une procédure séquentielle d'observation, sont construits sur le même principe à une distinction près : pour chaque valeur $\nu = 1, 2, \ldots, n, n+1, \ldots$ de la taille de l'échantillon, le domaine Γ_ν des valeurs possibles de la statistique critique $\gamma(X_1, \ldots, X_\nu)$ est partagé en trois régions disjointes : la région $\Gamma_\nu^{H_0}$ des valeurs vraisemblables, la région $\Gamma_\nu^{H_1}$ des valeurs invraisemblables, la région Γ_ν^* des valeurs douteuses (sachant que l'hypothèse H_0 est vraie), c'est-à-dire que

$$\Gamma_\nu = \Gamma_\nu^{H_0} \cup \Gamma_\nu^{H_1} \cup \Gamma_\nu^*, \quad \nu = 1, 2, \ldots$$

A chaque pas ν de la procédure séquentielle d'observation, c'est-à-dire après avoir enregistré les observations $X_1, \ldots, X_\nu$, $\nu = 1, 2, \ldots$, on prend une décision en se fondant sur la règle suivante :

si $\gamma(X_1, \ldots, X_\nu) \in \Gamma_\nu^{H_0}$ on accepte l'hypothèse H_0 :

si $\gamma(X_1, \ldots, X_\nu) \in \Gamma_\nu^{H_1}$ on rejette l'hypothèse H_0 (ou on accepte une hypothèse concurrente H_1) ;

si $\gamma(X_1, \ldots, X_\nu) \in \Gamma_\nu^*$ le problème reste ouvert jusqu'à la $(\nu+1)$-ième observation (pour cette raison, la région Γ_ν^* est appelée *région d'indétermination* ou encore *région de poursuite des observations*).

Donc, la concrétisation d'un test passe par la détermination : a) du type de l'hypothèse de base ; b) de la méthode de construction de la statistique critique $\gamma(X_1, \ldots, X_n)$; c) de la méthode de construction des régions $\Gamma_\nu^{H_0}$, $\Gamma_\nu^{H_1}$ et Γ_ν^* en fonction des caractéristiques données du test.

Comme exemple de test séquentiel, considérons le test du rapport de vraisemblance de Wald [140] entre deux hypothèses simples de la forme (9.13).

La statistique critique de ce test est définie pour la suite d'observations indépendantes $X_1, \ldots, X_\nu$ par la relation

$$\gamma^{(\nu)} = \ln \frac{f(X_1;\ \Theta_1) \ldots f(X_\nu;\ \Theta_1)}{f(X_1;\ \Theta_0) \ldots f(X_\nu;\ \Theta_0)} = \sum_{i=1}^{\nu} \ln \frac{f(X_i;\ \Theta_1)}{f(X_i;\ \Theta_0)}, \qquad \nu = 1, 2, \ldots \quad (9.15)$$

Les *régions des valeurs vraisemblables* ($\Gamma_\nu^{H_0}$), *invraisemblables* ($\Gamma_\nu^{H_1}$) et *douteuses* (Γ_ν^*) de la statistique critique $\gamma^{(\nu)}$, sachant que H_0 est

vraie, sont définies approximativement par les relations:

$$\Gamma_\nu^{H_0} = \left\{\gamma : \gamma^{(\nu)} \leqslant \ln \frac{\beta}{1-\alpha}\right\};$$

$$\Gamma_\nu^{H_1} = \left\{\gamma : \gamma^{(\nu)} \geqslant \ln \frac{1-\beta}{\alpha}\right\} \tag{9.16}$$

$$\Gamma_\nu^{*} = \left\{\gamma : \ln \frac{\beta}{1-\alpha} < \gamma^{(\nu)} < \ln \frac{1-\beta}{\alpha}\right\}.$$

Wald et Wolfowitz [140] ont prouvé l'optimalité de ce test séquentiel, plus exactement: de tous les tests entre les hypothèses (9.13) avec des risques de première et de deuxième espèce inférieurs à des valeurs respectives α et β données, le *test* (9.15), (9.16) *implique le plus petit nombre moyen d'observations* $\mathbf{E}_{\Theta_i}\,\nu\,(\alpha, \beta)$ que l'hypothèse H_0 $(i = 0)$ ou l'hypothèse H_1 $(i = 1)$ soit vraie. Wald a estimé [140] que son test était environ deux fois plus économique (quant au nombre d'observations) que le meilleur des tests classiques: le test de Neyman-Pearson. Mais, en 1959, S. Aïvazian a établi des formules asymptotiquement (par rapport au rapprochement indéfini des hypothèses testées) exactes pour $\mathbf{E}_{\Theta_i}\,\nu\,(\alpha, \beta)$ [5]:

$$\mathbf{E}_{\Theta_0}\nu\,(\alpha, \beta) = \frac{2\omega\,(\alpha, \beta)}{\rho\,(H_0, H_1)}; \quad \mathbf{E}_{\Theta_1}\nu\,(\alpha, \beta) = \frac{2\omega\,(\beta, \alpha)}{\rho\,(H_0, H_1)},$$

où $\omega\,(x, y) = (1 - x)\ln\frac{1-x}{y} + x\ln\frac{x}{1-y}$, et $\rho\,(H_0, H_1)$ est la « distance » entre les hypothèses testées (cf. § 9.4), qui ont permis, la formule (9.14) aidant, à comparer les bonnes propriétés des tests de Wald et de Neyman-Pearson:

$$\frac{\mathbf{E}_{\Theta_0}\nu\,(\alpha, \beta)}{n\,(\alpha, \beta; \rho)} = \frac{\omega\,(\alpha, \beta)}{(u_{1-\alpha}+u_{1-\beta})^2} = \tilde{\omega}\,(\alpha, \beta);$$

$$\frac{\mathbf{E}_{\Theta_1}\nu\,(\alpha, \beta)}{n\,(\alpha, \beta; \rho)} = \frac{\omega\,(\beta, \alpha)}{(u_{1-\alpha}+u_{1-\beta})^2} = \tilde{\omega}\,(\beta; \alpha).$$

Le tableau 9.1 représente les valeurs de la fonction $\tilde{\omega}\,(\alpha, \beta)$ correspondant aux risques de première et de deuxième espèce les plus souvent utilisés.

Sur ce tableau on voit que le gain en observations fourni par le test de Wald par rapport au test de Neyman-Pearson est de 2 à 3 fois supérieur et peut être bien plus élevé pour certaines combinaisons des risques (on démontre, en particulier que, $\lim\limits_{\alpha=\beta\to 0} \tilde{\omega}\,(\alpha, \beta) = \lim\limits_{\alpha=\beta\to 0} \tilde{\omega}\,(\beta, \alpha) = \frac{1}{4}$).

9.5.3. Test séquentiel généralisé entre hypothèses multiples. Le test de deux hypothèses multiples de la forme H_0: $\Theta \in \Delta_0$ et H_1: $\Theta \in \Delta_1$ où Θ est un paramètre (généralement, de dimension l) dont

Tableau 9.1

α \ β	0,2	0,1	0,05	0,02	0,01	0,005	0,002	0,001	0,0005	0,0002	0,0001
0,2	0,587	0,605	0,616	0,628	0,635	0,640	0,646	0,650	0,653	0,658	0,660
0,1	0,508	0,534	0,555	0,575	0,587	0,597	0,609	0,610	0,623	0,631	0,637
0,05	0,434	0,466	0,489	0,514	0,529	0,543	0,558	0,567	0,577	0,587	0,594
0,02	0,354	0,388	0,415	0,443	0,460	0,475	0,493	0,504	0,515	0,527	0,536
0,01	0,307	0,342	0,369	0,398	0,416	0,432	0,450	0,462	0,473	0,487	0,496
0,005	0,269	0,304	0,331	0,360	0,379	0,395	0,414	0,426	0,438	0,452	0,486
0,002	0,230	0,264	0,298	0,320	0,338	0,355	0,374	0,386	0,398	0,412	0,422
0,001	0,207	0,237	0,266	0,295	0,313	0,329	0,348	0,361	0,373	0,387	0,397
0,0005	0,188	0,220	0,245	0,277	0,291	0,308	0,326	0,339	0,350	0,365	0,375
0,0002	0,167	0,198	0,223	0,250	0,267	0,283	0,302	0,314	0,326	0,340	0,349
0,0001	0,154	0,184	0,208	0,235	0,252	0,282	0,286	0,298	0,309	0,323	0,333

dépend la loi de probabilité de la variable aléatoire étudiée, et Δ_0 et Δ_1 des régions disjointes de valeurs possibles de cette variable, se ramène à un test entre deux hypothèses simples de la forme (9.13), où Θ_0 et Θ_1 sont des points (« convenablement » choisis) respectivement de Δ_0 et Δ_1.

Mais dans cette position du problème, le test de Wald *ne jouit plus de ses bonnes propriétés*, car la vraie valeur du paramètre testé peut être égale à une valeur « intermédiaire » Θ^* (comprise entre Θ_0 et Θ_1), et le nombre moyen des observations $\mathbf{E}_{\Theta_i}\nu\,(\alpha,\ \beta)$ n'est minimal que si $\Theta = \Theta_0$ ou $\Theta = \Theta_1$.

Dans [6] on propose une méthode approchée de construction d'un test séquentiel généralisé optimal (TSGO) d'hypothèse de la forme (9.13) dans le cas où la vraie valeur du paramètre est susceptible d'être égale à Θ^*, où $\Theta^* \neq \Theta_0$ et $\Theta^* \neq \Theta_1$. L'optimalité de ce test traduit le fait que de tous les tests entre des hypothèses H_0 et H_1 de risques de première et de deuxième espèce respectivement inférieurs à des valeurs α et β données, il est celui qui donne lieu au plus petit nombre moyen d'observations dans la situation la « plus défavorable », c'est-à-dire à la plus petite valeur de $\max\limits_{\Theta} \mathbf{E}_\Theta\,\nu\,(\alpha,\ \beta)$. La statistique critique du TSGO est définie comme dans le test de Wald par la relation (9.15). Les régions d'acceptation de l'hypothèse H_0 (la région $\Gamma_\nu^{H_0}$), d'acceptation de l'hypothèse H_1 (la région $\Gamma_\nu^{H_1}$) et de poursuite des observations (la région Γ_ν^*) sont définies par les relations :

$$\Gamma_\nu^{H_0} = \left\{\gamma : \gamma^{(n)} \leqslant -c_0\left(1 - \frac{\nu\rho\,(H_0,\ H_1)}{8\ln\dfrac{1}{\min(\alpha,\ \beta)}}\right)\right\};$$

$$\Gamma_\nu^{H_1} = \left\{\gamma : \gamma^{(n)} \geqslant c_1\left(1 - \frac{\nu\rho\,(H_0,\ H_1)}{8\ln(1/\min(\alpha,\ \beta))}\right)\right\}; \qquad (9.17)$$

$$\Gamma_{\nu}^{*} = \left\{ \gamma : -c_0 \left(1 - \frac{\nu\rho(H_0, H_1)}{8 \ln (1/\min(\alpha, \beta))} \right) < \gamma^{(n)} < \right.$$
$$\left. < c_1 \left(1 - \frac{\nu\rho(H_0, H_1)}{8 (\ln (1/\min(\alpha, \beta))} \right) \right\},$$

où c_0 et c_1 sont des nombres strictement positifs approchés grossièrement respectivement par $2 \ln \frac{1}{\beta}$ et $2 \ln \frac{1}{\alpha}$.

Si l'on considère les domaines $\Gamma_{\nu}^{H_i}$ et Γ_{ν}^{*} dans le plan $(\nu, \gamma^{(\nu)})$, on constate que dans le test séquentiel de Wald leurs frontières sont définies par les droites $\gamma^{(n)} = -\ln \frac{1-\alpha}{\beta}$ et $\gamma^{(\nu)} = \ln \frac{1-\beta}{\alpha}$ qui sont *parallèles à l'axe horizontal*, tandis que dans le TSGO, leurs frontières sont deux droites concourrantes se coupant au point $\nu^* = \frac{1}{\rho(H_0, H_1)} \cdot 8 \ln \frac{1}{\min(\alpha, \beta)}$. Ceci exprime, en particulier, que le TSGO est un test séquentiel *tronqué*, c'est-à-dire tel que le nombre ν d'observations auquel il donne lieu est inférieur à un seuil ν^* donné. Le TSGO est décrit en détail dans [6].

Conclusions

1. La procédure de confrontation de l'hypothèse avancée par le chercheur sur la nature ou la valeur des paramètres inconnus du système stochastique envisagé, aux résultats des observations par tel ou tel test s'appelle test statistique de cette hypothèse.

2. Les hypothèses peuvent porter sur

la forme générale de la loi de probabilité de la variable aléatoire étudiée;

l'homogénéité de deux ou plusieurs échantillons;

les valeurs numériques des paramètres de la population générale étudiée;

la forme générale de l'interdépendance des composantes de la variable aléatoire multidimensionnelle étudiée;

l'indépendance et la stationnarité de la série d'observations.

3. Le principe de construction d'un test statistique est le suivant: a) on définit le type de l'hypothèse de base; b) on propose et on justifie la forme de la fonction d'observations (la statistique critique $\gamma^{(n)}$) sur le vu des valeurs de laquelle on prend la décision finale; c) on indique une méthode de définition, dans le domaine des valeurs possibles de la statistique critique $\gamma^{(n)}$, de la région $\Gamma_{n}^{H_1}$ de rejet de l'hypothèse de base H_0, telle que soit satisfaite la condition portant sur la valeur du risque de l'hypothèse H_0 (c'est-à-dire du niveau α du test).

4. La « qualité » d'un test se juge à son niveau α, sa puissance $1 - \beta$, sa propriété d'être sans biais et sa convergence. Si un test est

convergent, on peut rendre les erreurs de première et de deuxième espèce (α et β) aussi petites que l'on veut uniquement en faisant croître le nombre n des observations, sur le vu desquelles on prend la décision. Si le nombre n des observations est fixe, seule l'une des erreurs α ou β peut être rendue aussi petite que l'on veut, l'autre étant amené à croître inévitablement.

5. Lorsque la taille n de l'échantillon est fixée, on peut, en plus de la procédure classique d'observation, utiliser une procédure séquentielle, dans laquelle à chaque observation on prend l'une des trois décisions suivantes: « accepter l'hypothèse H_0 », « rejeter l'hypothèse H_0 », « ne pas prendre de décision et poursuivre les observations ». Ceci étant, la prise de décision dépend des résultats des observations précédentes, et le nombre ν des observations effectuées avant la prise de décision est une variable aléatoire.

6. Les tests séquentiels optimaux du rapport de vraisemblance (le test de Wald, le test séquentiel généralisé, etc.) sont les plus économiques quant au nombre d'observations nécessaires pour vérifier l'hypothèse de base avec une précision (α, β) donnée. Les recherches ont montré que les tests séquentiels permettent de réduire de deux, trois et même quatre fois le nombre d'observations nécessaires par rapport aux tests optimaux classiques.

QUATRIÈME PARTIE

TRAITEMENT STATISTIQUE PRIMAIRE DES DONNÉES

CHAPITRE 10

STATISTIQUE DESCRIPTIVE

10.1. Dossier de l'analyse ; introduction et mémorisation des données ; inspection des données

10.1.1. Dossier d'étude. Même si les recherches statistiques sont peu volumineuses et non itératives, les efforts dépensés pour une description exhaustive et opportune des fichiers utilisés et de toutes les étapes de l'analyse statistique sont rapidement compensés. Une préparation précoce et soignée du dossier d'étude évite de nombreux désagréments. Les grosses études statistiques sont conduites collectivement, la composition de l'équipe des chercheurs subit souvent des modifications, le traitement des données recueillies est échelonné dans le temps et implique des retours en arrière pour tester les hypothèses avancées en cours d'analyse. Dans de nombreuses recherches (notamment médicales) on assiste à une mise à jour permanente des données initiales. C'est dire qu'une organisation intelligente du dossier devient un élément capital pour un bon passage du « témoin » dans la course au résultat final. Attardons-nous succinctement sur certains aspects de ce processus.

Identification de l'analyse, des fichiers, des variables et des méthodes d'investigation. A chacun de ces éléments il est souhaitable d'associer : 1) un *identificateur* (un nom ou un matricule) *abrégé* qui apparaîtra à chaque sortie ; 2) un *identificateur complet* qui figurera essentiellement dans les comptes rendus, mais que l'on retrouvera parfois dans les fichiers de sortie lorsque l'identificateur abrégé risquera de prêter à équivoque ; 3) un *plan* succinct de travail indiquant les liens entre les divers fichiers, les conditions de recueil ou d'élaboration de ces fichiers, une méthode d'obtention, de mesure ou d'enregistrement des variables, les sources des méthodes d'investigation. Les descriptions sont utilisées essentiellement pour l'élaboration des comptes rendus et parfois dans les commentaires destinés à faciliter l'intelligence de certaines sorties ; 4) l'*indication* pour les variables seulement *de l'intervalle de variation* ou des *valeurs prises*, cette information devant nécessairement être utilisée à l'entrée pour

le contrôle des données et à la sortie, pour l'élaboration des tableaux de sortie.

Si en cours d'analyse certains fichiers sont privilégiés ou si de nouvelles variables sont introduites, il est nécessaire de les décrire avec le même soin que les principaux fichiers et variables.

A l'étape actuelle du développement du logiciel, l'automatisation de la documentation est réalisée par des procédés relativement simples et permet de résoudre de très importants problèmes, tels que le contrôle des variables d'entrée; la « lisibilité autonome » de tous les tableaux délivrés; un meilleur dépistage des erreurs de description, l'élaboration plus aisée des comptes rendus.

Par ailleurs, il serait souhaitable d'introduire dans l'ordinateur ou dans des cartes spéciales: les formes d'analyse (les programmes) et les sous-fichiers auxquels elles ont été appliquées; le degré de dépendance des diverses variables, de validité de la prédiction, de pertinence de la projection des objets dans un espace de moindre dimension, etc.; les adresses des sorties respectives, ainsi que des commentaires textuels concernant aussi bien la logique et le cours de l'analyse que les divers listings.

10.1.2. Introduction et stockage des données. L'introduction des données s'opère soit par des cartes perforées, soit par visualisation des variables codées, soit enfin par visualisation de la liste des valeurs possibles de la variable, ou comme disent les informaticiens, par visualisation du « menu ». Les deux derniers procédés permettent de détecter momentanément les fautes graves introduites. L'introduction du menu demande beaucoup de temps. Le menu doit décrire automatiquement les variables. Les données doivent être stockées de façon à être facilement corrigeables et complétables.

10.1.3. Inspection des données. Il est très important de bien rédiger les données recueillies et de les passer au peigne fin avant de les soumettre au principal traitement statistique. Les erreurs affectant les données peuvent conduire à des résultats insolites parfois interprétables, parfois non, mais invariablement faux.

L'inspection des données a pour but

1) de dépister les erreurs grossières dans le dictionnaire des variables, ainsi que les erreurs commises lors du codage, de la perforation des cartes et de l'introduction des données dans l'ordinateur;

2) de signaler les éventuels dépassements ou observations anomales qui risquent de ne pas être représentatives de la population étudiée (pour plus de détails, voir § 11.5);

3) de donner une idée approximative des lois à une dimension et partiellement à deux dimensions.

Voici quelques méthodes destinées à faciliter l'inspection des données.

Le *listing des données* introduites dans l'ordinateur sous forme

d'un tableau des objets rangés, parfois selon la valeur d'un caractère. Ce faisant, on s'assure que des erreurs grossières n'ont pas été commises lors de la définition de l'échelle des données, que l'identification de l'analyse et des variables est correcte et facilement lisible, que les données sont exhaustives et non redondantes et, enfin, que les valeurs numériques des variables ou de leurs codes tombent bien dans l'intervalle prévu. L'inspection de variables rangées par colonnes permet de localiser immédiatement les erreurs grossières. Si besoin les colonnes peuvent être visualisées sur un écran. Mais un listing bien rédigé est un important document de référence pour répondre aux questions susceptibles de se poser aux étapes ultérieures de l'analyse.

Construction des lois à une dimension. Si l'ordinateur construit l'histogramme (cf. § 10.3), il est commode de remplir ses colonnes avec les numéros des observations. Si les observations sont trop nombreuses, il faut mentionner à part les numéros des observations qui ont dépassé les quantiles d'ordre 0,05 et d'ordre 0,95.

Il est commode d'indiquer les numéros des observations sur les listings à deux dimensions. Si plusieurs observations tombent en un même point, on note cette coïncidence par un signe spécial sous lequel on imprimera les numéros de ces observations. Les listings à deux dimensions sont précieux pour l'établissement de conjectures préliminaires sur les relations entre les variables.

Les problèmes mathématiques soulevés par la construction des fonctions de répartition empiriques sont traités au § 10.3.

10.2. Echelles de mesures

Chaque mesure d'un objet est réalisée dans une certaine échelle. Les coordonnées d'un vecteur d'observations peuvent être exprimées dans des échelles différentes. Au § 5.1 on a cité un exemple de vecteur d'observations (tableau 5.1) dont les premières coordonnées expriment un état, une qualité (milieu social de la famille, sexe et profession du chef de famille, qualité des conditions de logement), alors que les autres sont des nombres (le nombre des membres de la famille, le nombre d'enfants, le revenu moyen annuel, etc.). Les propriétés de ces échelles diffèrent fortement. Ainsi, à propos du sexe du chef de famille, on peut dire qu'il est ou masculin ou féminin et que le sexe masculin est différent du sexe féminin ; à propos des conditions de logement, qu'elles sont ou ne sont pas les mêmes et que certaines conditions sont meilleures que d'autres ; à propos des dépenses, qu'une famille dépense plus, moins ou la même somme pour son alimentation qu'une autre, on peut faire la différence des dépenses, le rapport des dépenses.

On décrit plus bas les principaux types d'échelles et les méthodes mathématiques d'unification des données mesurées dans des échelles différentes.

10.2.1. Echelle nominale. Cette échelle n'est utilisée que pour ranger un individu ou un objet dans une classe déterminée. Si les classes et les règles de classification sont définies *a priori*, on parle d'une *échelle catégorisée*, sinon, d'une *échelle non catégorisée*. Le sexe est un exemple d'échelle catégorisée : à chaque individu on associe la lettre M ou la lettre F, le nombre 1 ou le nombre 2 ou encore un signe spécial. En fait, on aurait pu utiliser d'autres lettres et chiffres, l'essentiel est que la correspondance entre les codes soit biunivoque. Pour introduire les données catégorisées, il est commode d'utiliser le « menu », c'est-à-dire la liste des éventuelles catégories avec leurs codes. Le nom, le prénom, la date de naissance sont des exemples de variables nominales non catégorisées.

Une autre source importante de variables nominales non catégorisées est indiquée au § 5.3. C'est le cas où l'observation porte sur un couple d'objets et la variable indique seulement si ces objets appartiennent ou non à une même classe sans spécifier laquelle. Cette situation n'a rien d'étonnant. Certes, si les classes sont définies à l'avance et si la classification ne pose pas de problème, il faut indiquer à quelle classe appartient l'objet étudié. Mais parfois ces classes ne sont pas données *a priori* et la classification des objets est le but de l'analyse. On peut, par exemple, parler de l'évolution « proche » ou « semblable » d'une maladie chez deux personnes, bien que l'on ne connaisse pas intégralement le tableau clinique de cette maladie. Bien plus, la connaissance de formes voisines d'une maladie peut servir de point de départ à la description du processus morbide. Ceci est valable aussi pour les groupes socio-économiques, etc.

Une même variable peut remplir des fonctions différentes. Par exemple, la variable nominale non catégorisée : le nom du programme, ne sert qu'à identifier un programme et peut être trouvée par une simple inspection de la liste de tous les programmes. Si les programmes sont classés d'une manière ou d'une autre (par exemple, par ordre alphanumérique), le nom d'un programme revêt le caractère d'une variable ordinale. A propos de deux noms, on dira qu'ils sont confondus ou que l'un précède l'autre dans cette méthode d'ordonnancement. L'ordre de succession change avec la méthode d'ordonnancement.

Les opérations arithmétiques sur les variables nominales sont dénuées de sens. Donc, ni la *médiane* ni la *moyenne arithmétique* ne peuvent être utilisées comme mesure cohérente de la tendance centrale. Le *mode* convient mieux en l'occurrence.

10.2.2. Echelle ordinale. Dans l'échelle ordinale, les classes sont rangées selon le degré d'expression d'un caractère donné. A cha-

que classe on associe un symbole de telle sorte que l'ordre des symboles et celui des classes se correspondent. Les classes seront rangées dans l'ordre de croissance numérique si on leur associe des nombres; dans l'ordre alphabétique, si on leur associe des lettres; selon la valeur des mots, si on leur associe des mots. Au § 5.3 on cite un exemple d'échelle ordinale pour la description des conditions de logement à quatre modalités (classes): « mauvaises », « satisfaisantes », « bonnes ». « très bonnes ». On aurait pu très bien numéroter ces classes avec les nombres 1, 2, 3, 4 ou 4, 3, 2, 1 ou leur associer les lettres a, b, c, d, etc.

Comme autres exemples d'échelles ordinales citons: en médecine, l'échelle Miasnikov des divers degrés de l'hypertension, l'échelle Strajesko-Vassilenko-Lang des formes de l'insuffisance cardiaque, l'échelle Fogelson de l'insuffisance coronarienne; en minéralogie, l'échelle Moos qui classe les minéraux d'après leur dureté en attribuant 1 au talc, 2 au gypse, 3 à la calcite, 4 à la fluorine, 5 à l'apatite, 6 à l'orthose, 7 au quartz, 8 à la topaze, 9 au corindon et 10 au diamant; en géographie, l'échelle de Beaufort qui classe les vents en « calme », « faible », « doux », etc.

La structure d'une échelle ordinale n'est pas modifiée par les transformations biunivoques des codes qui préservent l'ordre. Comme dans les échelles nominales, les opérations arithmétiques perdent leur sens par une transformation des échelles ordinales, donc il est préférable de ne pas s'en servir. On démontre sans peine que si l'on raisonne uniquement en termes d'échelles, les seules statistiques autorisées, quand on a affaire à des échelles ordinales, sont les termes d'une série variationnelle [107].

10.2.3. Echelles quantitatives. On appelle échelle d'intervalles une échelle qui exprime *de combien* un objet se distingue d'un autre par le degré d'expression d'un caractère donné. Pour se donner une échelle d'intervalles, il faut déterminer les objets correspondant au *point initial* et l'*unité de mesure*, et, ensuite, associer à chaque objet un nombre indiquant de combien d'unités de mesure cet objet se distingue de l'objet de référence. Un exemple classique d'échelle d'intervalles est l'échelle Celsius, où 0° est pris pour point initial et où l'unité de mesure est 1°.

La structure d'une échelle d'intervalles est invariante par les applications linéaires de la forme $y = ax + b$, $a > 0$. Cette application déplace l'origine de b unités et multiplie l'unité de mesure par a.

Par exemple, l'application $y = \frac{9}{5}x + 32$, où x est la température en degré Celsius, permet de passer de l'échelle Celsius à l'échelle Fahrenheit.

Si l'origine de l'échelle d'intervalles est le zéro absolu, on a la possibilité d'exprimer *combien* de fois une mesure est contenue dans une autre. Une telle échelle s'appelle *échelle des rapports*. L'échelle des rapports est justiciable de transformations de la forme $y = ax$, $a > 0$. La plupart des échelles utilisées en physique sont soit des échelles d'intervalles (pour mesurer les températures, l'énergie potentielle, etc.), soit des échelles des rapports (pour mesurer le temps, la masse d'un corps, la charge, les distances).

Vu que les échelles quantitatives sont justiciables de transformations arithmétiques, la moyenne arithmétique peut être utilisée pour agréger les données.

10.2.4. Unification des données. A chaque type d'échelle est associée une technique de traitement statistique. Ainsi, pour les variables mesurées dans une échelle nominale on peut se servir du test du χ^2 pour les lois polynomiales, du test du χ^2, pour s'assurer de l'absence d'associations dans les tableaux de contingence. du test des hypothèses concernant les valeurs du paramètre de la loi binomiale. Si l'échelle est ordinale, on applique des méthodes utilisant les rangs (corrélation des rangs, tests non paramétriques d'hypothèses de la forme $F(x) \leqslant G(x)$, etc.). Dans le cas d'une échelle d'intervalles, on peut mettre en œuvre l'arsenal tout entier des méthodes statistiques.

Par ailleurs, on a élaboré des procédures statistiques pour les cas où les coordonnées des vecteurs observés sont mesurées dans des échelles différentes. Un exemple typique est l'analyse de variance (cf. § 3.5), dans laquelle les facteurs sont mesurés dans une échelle nominale et les sorties correspondant à leurs combinaisons, dans une échelle d'intervalles.

Cependant dans de nombreuses méthodes statistiques et notamment dans les méthodes modernes de l'analyse multidimensionnelle, on admet que les données sont mesurées dans des échelles unifiées. Diverses méthodes d'unification des données ont été proposées. Voyons les plus importantes d'entre elles.

Réduction à des variables binaires. Cette méthode est basée sur la substitution à chaque aléatoire primitive d'une série de variables aléatoires prenant seulement les valeurs 0 et 1. Pour une variable nominale x à k modalités $x_1, \ldots, x_k$, on introduit k variables $y_1, \ldots, y_k$, telles que $y_i = 1$ lorsque $x = x_i$ et $y_i = 0$ lorsque $x \neq x_i$ $(i = 1, \ldots, k)$.

Ce procédé est parfois utilisé pour réduire une variable aléatoire mesurée dans une échelle ordinale à des variables binaires. Mais souvent il est plus commode de distinguer l'événement $x \geqslant x_k$ au lieu de $x = x_k$. Comparons les qualités de ces méthodes à la faveur du problème suivant. Soient ξ une variable aléatoire uniformément

répartie sur l'intervalle [0, 1], δ un petit nombre;

$$\delta/2 \leqslant x \leqslant y - \delta < 1 - 3\delta/2;$$

$$\begin{cases} u = u(x) = 1 & \text{si } x - \delta/2 \leqslant \xi < x + \delta/2; \\ u(x) = 0 & \text{sinon}; \end{cases}$$

$$\begin{cases} v = v(x) = 1 & \text{si } \xi \geqslant x; \\ v(x) = 0 & \text{si } \xi < x. \end{cases}$$

La fonction $u(x)$ modélise visiblement la première méthode de passage aux variables binaires, la fonction $v(x)$, la seconde. Des calculs peu compliqués nous donnent:

$$\operatorname{cor}(\xi, u(x)) = (x - 0{,}5)\sqrt{\delta} \cdot 12/\sqrt{1-\delta}; \tag{10.1}$$

$$\operatorname{cor}(u(x), u(y)) = -\delta/(1-\delta); \tag{10.2}$$

$$\operatorname{cor}(\xi, v(x)) = 0{,}5\sqrt{x(1-x)} \cdot \sqrt{12}; \tag{10.3}$$

$$\operatorname{cor}(v(x), v(y)) = \sqrt{x(1-y)}/\sqrt{(1-x)y}. \tag{10.4}$$

Des formules (10.1) et (10.2) il s'ensuit que la corrélation entre ξ et $u(x)$ dépend de la variable x. Le coefficient de corrélation est de même signe dans la première et dans la seconde moitié de l'intervalle [0, 1] et s'annule pour $x = 0{,}5$. La corrélation entre $u(x)$ et $u(y)$ ne dépend pas de la position relative de x et y. Dans la deuxième méthode de réduction aux variables binaires, le signe de la corrélation entre ξ et $v(x)$ est le même sur l'intervalle tout entier et la corrélation, la plus grande pour $x = 0{,}5$ (formule (10.3)). La corrélation entre $v(x)$ et $v(y)$ est toujours strictement positive et d'autant plus élevée que x est proche de y (formule (10.4)), ce qui est légitime. Donc, dans le deuxième procédé de réduction aux variables binaires, les relations de corrélation gardent en grande partie leur ancienne signification.

Si l'on a affaire à une variable quantitative continue, on quantifie ses valeurs, on les regroupe en modalités, puis, on procède comme pour l'échelle ordinale.

Le principal défaut de la technique exposée est l'introduction d'un grand nombre de nouvelles variables et une perte partielle d'information contenue dans les données à cause aussi bien de la quantification que de la réduction artificielle du niveau de l'échelle utilisée.

Codage des variables ordinales et nominales. Cette méthode est à l'antipode de la méthode qui vient d'être exposée: en effet, toutes les variables sont hissées au rang de variables quantitatives par attribution de valeurs numériques à leurs modalités. Les valeurs affectées sont parfois appelées codes.

Le choix des codes dépend essentiellement des objectifs fixés au codage. Si, par exemple, l'on étudie le lien entre deux variables nominales, il faut choisir les codes à partir de la condition de maximisation du coefficient de corrélation de ces variables [50], [85]. S'il est question d'une classification des observations (analyse discriminante), le choix des codes peut être rattaché à la condition de maximisation de la distance normée dans un espace des échantillons à plusieurs dimensions entre les centres des populations étudiées (la distance de Mahalanobis). On simplifie parfois ce problème en affectant ces codes aux coordonnées de façon à maximiser seulement la distance normée entre les valeurs moyennes de la coordonnée envisagée. Dans [98] et [100] on compare statistiquement sur un exemple particulier l'efficacité de ces deux méthodes de codage en analyse discriminante.

Les méthodes de codage, dans lesquelles les codes sont choisies à partir de la condition de maximisation d'une fonctionnelle convenablement définie, s'inscrivent dans le cadre de l'approche extrémale de la formulation des principaux problèmes de mathématique statistique mentionnée au § 1.2.

En général, le codage des variables qualitatives est un problème compliqué tant sur le plan numérique que sur le plan statistique. Certains aspects de ce problème sont discutés dans [48, 50, 77, 82, 85].

10.3. Lois empiriques

A l'état brut les données sont généralement peu « locaces ». Avant toute analyse il faut les ranger et les apprêter. En particulier, il serait souhaitable d'avoir une idée de la loi de probabilité des variables aléatoires figurant dans ces données.

10.3.1. Histogramme. Si le nombre des valeurs possibles de la variable aléatoire n'est pas élevé, on peut se faire une idée de sa loi de probabilité en étudiant les fréquences d'apparition de chacune de ses valeurs.

Dans le cas général, on regroupe les valeurs des variables aléatoires en classes et on compte combien de fois ces valeurs apparaissent dans chaque classe. Au lieu des innombrables valeurs éparpillées, on obtient un tableau statistique assez suggestif. La technique de passage aux données groupées est développée en détail au § 5.4.

Pour représenter la loi de probabilité de façon plus suggestive, on construit dans un système de coordonnées rectangulaires une figure spéciale appelée histogramme de la loi. A cet effet, on partage l'axe horizontal en intervalles égaux correspondant aux classes et sur chacun de ces intervalles pris pour base on construit un rectangle dont la hauteur est proportionnelle à la fréquence de la classe respective. L'histogramme obtenu dépend du choix de la longueur

des intervalles de classes. Pour réduire cette dépendance, on procède à un lissage de ces histogrammes. Une méthode de lissage consiste à relier par des droites les milieux des bases supérieures des rectangles voisins.

Exemple 10.1. Le tableau 10.1 représente les valeurs du logarithme u du salaire (en unités conventionnelles) d'ouvriers à la pièce.

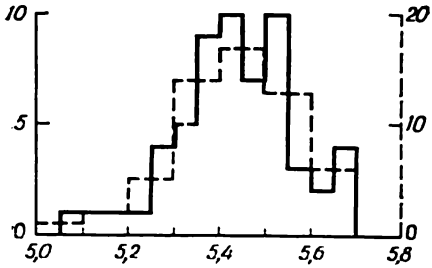


Fig. 10.1. Histogrammes rectangulaires de la répartition des données du tableau 10.1 pour deux intervalles de groupement (le nombre d'observations est porté sur l'axe des ordonnées, u, sur l'axe des abscisses)

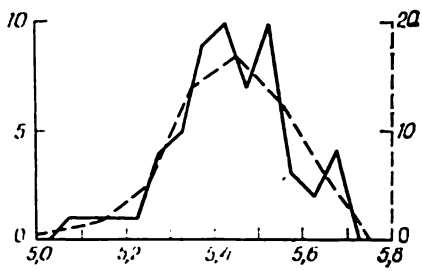


Fig. 10.2. Histogramme de la figure 10.1 après lissage

On sait que cette variable est approximativement normale. Pour illustrer l'influence de la longueur des intervalles de groupement sur la forme de l'histogramme, on a représenté sur la figure 10.1 des histogrammes de pas $h = 0{,}05$ et $h = 0{,}10$. On constate que l'histo-

Tableau 10.1

5,44	5,56	5,34	5,54	5,43	5,40
5,37	5,63	5,45	5,32	5,30	5,38
5,36	5,51	5,37	5,51	5,30	5,65
5,42	5,53	5,41	5,31	5,48	5,36
5,50	5,61	5,48	5,46	5,67	5,53
5,68	5,52	5,43	5,50	5,37	5,54
5,33	5,46	5,09	5,39	5,27	5,58
5,25	5,40	5,69	5,60	5,43	5,48
5,52	5,15	5,34	5,66	5,43	
5,52	5,41	5,17	5,44	5,39	

gramme de plus grand pas (la ligne en pointillé) est plus lisse. On observe le même effet sur la figure 10.2 qui représente les histogrammes lissés.

Les données groupées sont parfois utilisées pour calculer les moments de la variable aléatoire à la place des vraies valeurs des observations. Ceci étant, si les intervales de classes sont tous de même

longueur et si leur position est choisie de façon aléatoire, la valeur moyenne de l'échantillon calculée sur les données groupées, ne varie pas en moyenne, et la variance augmente de $h^2/12$ [102, 138]. Pour compenser cet effet, de la valeur de la variance calculée sur les données groupées on retranche la quantité $h^2/12$ (la correction de Sheppard). En particulier, pour $h = \sigma/2$, l'influence du groupement sur l'estimation de σ est de l'ordre de 1 % et on peut la négliger.

Exemple 10.2. Le tableau 10.2 représente les estimations de la moyenne et de la variance calculées sur les données groupées (fig. 10.1) et les données non groupées (tableau 10.1).

On remarque sur le tableau 10.2 que les valeurs moyennes varient peu, tandis que la variance croît mais est compensée dans une grande mesure par la correction de Sheppard.

Tableau 10.2

Données	$\bar{u}$	$s^2-h^2/12$	Correction de Sheppard
Initiales	5,439	0,01633	0
Premier groupement	5,434	0,01681	0,00021
Deuxième groupement	5,434	0,01703	0,00083

10.3.2. Estimations non paramétriques de la densité. Pour estimer la densité de probabilité, on peut outre les histogrammes se servir de la statistique

$$\hat{f}(x) = (bn)^{-1} \sum_{i=1}^{n} k((x_i - x)/b), \tag{10.5}$$

où b est un petit paramètre, $k(u)$ une fonction vérifiant les conditions suivantes : $k(u) \geqslant 0$, $k(u) = k(-u)$, $\int k(u)\,du = 1$, $k(u) \to 0$ ($|u| \to \infty$). Pour $k(u)$, on prend souvent la densité de la loi normale réduite. Le principal avantage de ces estimations sur les histogrammes est qu'elles ne dépendent pas du choix de la position des intervalles de classes. Cette circonstance facilite leur généralisation au cas multidimensionnel. En vérité, le choix du paramètre b est arbitraire tout comme la longueur de l'intervalle de classe.

Pour se faire une idée du mécanisme de fonctionnement des estimations (10.5), on a construit sur la figure 10.3 les graphiques de trois d'entre elles pour des valeurs différentes de b en utilisant les données du tableau 10.1. Comparez ce dessin aux histogrammes des figures 10.1 et 10.2.

10.3.3. Estimation de la fonction de répartition. Soit donné un échantillon $x_1, \ldots, x_n$ d'une population générale de fonction de

répartition $F(t)$. Pour estimation de $F(t)$, on prend

$$\hat{F}(t) = \frac{1}{n+1}\left(\frac{1}{2} + \sum_{x_i < t} 1\right) \tag{10.6}$$

où la sommation est étendue à tous les $x_i < t$. Le théorème limite central (cf. § 7.3) nous dit que pour tous t et $\varepsilon > 0$, on a

$$P\{|\hat{F}(t) - F(t)| > \varepsilon\} \to 0 \text{ lorsque } n \to \infty.$$

On obtiendrait une estimation voisine de $F(t)$ en intégrant l'estimation non paramétrique (10.5) de la densité entre $-\infty$ et t.

S'il faut vérifier qu'une variable aléatoire admet une fonction de répartition $G(t)$ appartenant à une famille $F((t-\mu)/\sigma)$,

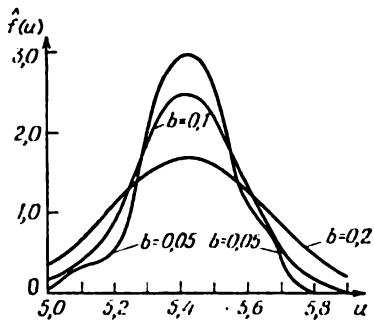


Fig. 10.3. Estimations non paramétriques de la densité de u pour trois valeurs de b

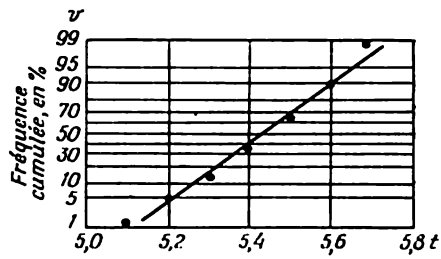


Fig. 10.4. Fonction de répartition de u sur du papier gausso-arithmétique

où $F(\cdot)$ est une fonction de répartition continue connue, on peut, pour construire $\hat{F}$, se servir d'une échelle spéciale qui consiste à porter sur l'axe des ordonnées non pas $\hat{F}(t)$ mais les valeurs de la fonction $v = F^{-1}(\hat{F}(t))$, où F^{-1} est l'inverse de F. Dans les nouvelles coordonnées (t, v), le graphique de $G(t)$ est une droite qui permet d'estimer sans peine les paramètres μ et σ (cf. n° 10.4.3). Il existe un papier spécial dit gausso-arithmétique dans lequel $v = \Phi^{-1}$, où $\Phi(\cdot)$ est la fonction de répartition de la loi normale réduite. Sur la figure 10.4 on a représenté sur ce papier la fonction de répartition empirique des données du tableau 10.1. Pour $n = 58$, on constate un alignement des points, ce qui exprime que la répartition de u est approximativement normale.

10.3.4. Transformation des variables. Presque toutes les méthodes de l'analyse multidimensionnelle partent d'une manière ou d'une autre de l'hypothèse de normalité des variables aléatoires figurant

dans le modèle. Ceci nous suggère de trouver une transformation qui rende les variables aléatoires non normales (non gaussiennes) approximativement normales. Des considérations géométriques suggèrent que toute répartition continue peut être rendue normale par une transformation monotone continue de la variable aléatoire. Initialement, on s'est servi des transformations $y = \sqrt{x}$ et $y = \log(x+a)$. Dans la monographie de Hald [68] on trouvera de nombreux exemples mettant en évidence l'interêt de ces transformations.

J. Box et D. Cox [31] recommandent d'utiliser les familles de transformations à un et deux paramètres suivantes :

$$y = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & (\lambda \neq 0)\,; \\ \ln x & (\lambda = 0)\,; \end{cases} \tag{10.7}$$

$$y = \begin{cases} \dfrac{(x+\lambda_2)^{\lambda_1}-1}{\lambda_1} & (\lambda_1 \neq 0)\,; \\ \ln(x+\lambda_2) & (\lambda_1 = 0). \end{cases} \tag{10.8}$$

Il faut prêter une attention toute particulière au lien entre les densités de probabilité $f(y)$ et $p(x)$ de y et de x (cf. § 7.4) :

$$p(x) = f(y)\left|\frac{\partial y}{\partial x}\right|.$$

Dans le cas de la famille à un paramètre (10.7),

$$\ln\frac{\partial y}{\partial x} = (\lambda-1)\ln x\,; \tag{10.9}$$

dans celui de la famille à deux paramètres (10.8),

$$\ln\frac{\partial y}{\partial x} = (\lambda_1-1)\ln(x+\lambda_2). \tag{10.10}$$

Les valeurs convenables du ou des paramètres peuvent être déterminées graphiquement, comme le préconise [68], ou par la méthode du maximum de vraisemblance en traitant λ (resp. λ_1, λ_2) comme le paramètre (resp. les paramètres) de la répartition et en écrivant la fonction de vraisemblance pour les variables primitives.

10.3.5. Tableaux de contingence. Les répartitions empiriques multidimensionnelles et notamment les répartitions des variables aléatoires nominales et ordinales sont souvent représentées par des tableaux rectangulaires avec une double, triple ou multiple classification. Dans la case correspondant à la i_1-ième modalité de la première classification, à la i_2-ième modalité de la deuxième classification, ..., à la i_l-ième modalité de la l-ième classification, on porte le nombre $x_{i_1 i_2 \ldots i_l}$ d'objets possédant simultanément les modalités respectives de chaque variable. Ces tableaux sont appelés

tableaux de contingence. De nombreux traités de statistique leur sont consacrés. Les techniques traditionnelles de traitement des tableaux de contingence sont accessibles dans [40, 79, 138], les techniques modernes utilisant les modèles loglinéaires, dans [15, 22, 66, 111].

10.4. Estimation des paramètres de localisation et d'échelle

10.4.1. Position du problème. Dans la pratique, on est souvent confronté à des répartitions proches des répartitions normales quand on enregistre directement les variables observées ou quand on transforme les données empiriques comme indiqué à la fin du § 10.3. Les principales caractéristiques que l'on utilise lorsqu'on manipule ces répartitions sont d'une manière ou d'une autre les paramètres de localisation et d'échelle. Vu que la forme analytique exacte des répartitions empiriques est généralement inconnue, et souvent de peu d'intérêt pour le chercheur, ces paramètres sont déterminés par une procédure numérique. Ceci étant, les estimations des paramètres doivent être telles que

1) si la répartition est normale, le paramètre de localisation empirique m soit une estimation convergente de la moyenne μ, et le paramètre d'échelle empirique b, une estimation convergente de l'écart-type σ;

2) les paramètres changent comme les paramètres d'une loi normale par une transformation linéaire de la variable aléatoire.

Dans le cas multidimensionnel, on introduit respectivement le vecteur de localisation M et une matrice semi-définie positive $\mathbf{B}$ qui se ramènent, pour une répartition normale, aux habituels vecteur des moyennes et matrice des covariances.

Chaque méthode d'estimation des paramètres de localisation et d'échelle conduit à une certaine paramétrisation des répartitions. Penchons-nous sur les diverses méthodes d'estimation. Ces 10 à 15 dernières années ces questions monopolisent de plus en plus l'attention des statisticiens de métier, puisqu'une modélisation mathématique directe a montré que les procédures statistiques, optimales pour la loi normale, perdent brusquement leurs propriétés optimales en cas d'alourdissement des ailes. Avant de passer à un exposé systématique de cette question, arrêtons-nous sur les éventuels mécanismes d'apparition des ailes « lourdes ». Un échantillon normal peut être pollué par des erreurs importantes d'ordre technique, commises lors de l'enregistrement des observations. Ces « aberrations » doivent être identifiées, puis, tout simplement abandonnées. Dans le deuxième cas qui, en principe, est important, nous avons affaire à une répartition approximativement normale mélangée à une autre de moyenne légèrement différente et de variance plus élevée. Ces cas se présentent lors de la définition des « normes » en économie, dans les études sociologiques, en médecine. Ici, il n'est plus possible

de traiter la « pollution » comme une « aberration », car les données enregistrées dans une partie de l'échantillon diffèrent parfois moins par les limites de variation des valeurs possibles des variables que par la variation de la fréquence des écarts par rapport à la tendance centrale de la « norme ». La situation est davantage compliquée par le fait qu'ici les pollutions non symétriques sont chose courante. Ainsi, par exemple, quand on estime le rendement du travail dans un secteur industriel, les entreprises fonctionnant sur des technologies dépassées auront en moyenne de moins bons indices que le principal groupe d'entreprises. De façon analogue, un début de maladie déplace les indices physiologiques du patient dans une seule direction. Le choix du procédé de paramétrisation dépend essentiellement dans ces cas de l'objectif de la recherche : faut-il décrire la population tout entière ou seulement son noyau (qui représente la « norme ») après avoir si possible éliminé l'effet de « pollution »? Aussi avant d'introduire les paramètres d'échelle et de localisation par tel ou tel procédé, doit-on veiller à la simplicité de l'interprétation des estimations, à la possibilité de les appliquer à la résolution des problèmes envisagés et aux propriétés statistiques empiriques non seulement pour la loi normale, mais aussi pour un large spectre de modèles de « pollution » de cette loi. Enfin, la simplicité de la réalisation technique des estimations, notamment dans le cas multidimensionnel, joue un rôle assez important.

10.4.2. Estimation des paramètres de la loi normale. Soit $X_1, \ldots, X_n$ un échantillon de taille n d'une répartition normale multidimensionnelle de paramètres (M, Σ). Les estimations sans biais et efficaces (cf. § 8.1) des paramètres sont définies par les formules

$$\bar{X} = \sum_{i=1}^{n} X_i/n; \qquad S = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'/(n-1) \text{ *)}. \tag{10.11}$$

La matrice des covariances de l'estimation de la moyenne est donnée par la formule

$$E(\bar{X} - M)(\bar{X} - M)' = n^{-1}\Sigma. \tag{10.12}$$

Les estimations $(\bar{X}, S)$ sont actuellement les plus sollicitées et figurent dans tous les paquets de programmes statistiques. Malheureusement, elles sont douées de mauvaises propriétés empiriques pour les répartitions voisines des normales mais possédant de plus lourdes ailes (les détails sont développés au n° 10.4.4).

10.4.3. Méthode graphique d'estimation. Traçons sur du papier gausso-arithmétique une droite correspondant à une fonction de

*) Si l'on respectait la logique de cette notation, on devrait désigner l'estimation quadratique moyenne de la variance de x par s en dimension un. Mais on se conformera à l'usage en statistique mathématique et on la notera s^2.

répartition (cf. n° 10.3.3) et soit u_α $(0 < \alpha < 1)$ l'abscisse correspondant à la probabilité α. Posons

$$\hat{\mu} = u_{1/2}, \quad \hat{\sigma}_{3/4,\,1/4} = (u_{3/4} - u_{1/4})/1{,}34$$

ou

$$\hat{\sigma}_{0{,}84;\,0{,}16} = (u_{0.84} - u_{0.16})/2.$$

Dans cette définition, $\hat{\mu}$ est la médiane de la répartition, le numérateur de $\hat{\sigma}_{3/4,\,1/4}$ est la *dispersion interquantile*. Dans le cas normal, ces estimations admettent les variances asymptotiques $(n \to \infty)$ suivantes [41]:

$$V\hat{\mu} \approx \frac{\pi}{2} \cdot \frac{\sigma^2}{n} \approx 1{,}57\frac{\sigma^2}{n}; \quad V\hat{\sigma}_{3/4,\,1/4} \approx 1{,}36 \cdot \frac{\sigma^2}{n};$$

$$V\hat{\sigma}_{0{,}84;\,0{,}16} \approx 0{,}93\frac{\sigma^2}{n}.$$

Si l'on tient compte du fait que, dans le cas d'une répartition normale, $V\bar{x} = \sigma^2/n$ et $V\bar{s} = \sigma^2/(2n)$, on constate que ces estimations sont moins efficaces que celles utilisées ordinairement pour le cas normal.

§ 10.4.4. Stabilité des estimations pour de petits écarts de la répartition par rapport à la normale. Gauss a débouché sur la loi normale en 1809 en étudiant un problème que dans le langage actuel on pourrait énoncer comme suit [57, 58]: trouver une répartition pour laquelle la moyenne arithmétique $\bar{x}$ des observations indépendantes serait l'estimation du paramètre de localisation par le maximum de vraisemblance. De la position même du problème de Gauss il appert que $\bar{x}$ n'est pas tenu de jouir de bonnes propriétés empiriques pour les répartitions qui ne sont pas normales. Et c'est effectivement le cas: l'efficacité de $\bar{x}$ en tant qu'estimation du paramètre de localisation baisse rapidement lorsque les ailes s'alourdissent même si les répartitions sont symétriques. Le tableau 8.1 représente l'efficacité relative de $\bar{x}$ pour de nombreuses répartitions.

Historiquement, on a remarqué depuis longtemps que, quand on estimait le centre d'une répartition, on avait intérêt à éliminer un même nombre peu élevé de termes extrêmes de la série variationnelle et à prendre la moyenne arithmétique des termes restants. Cette estimation était utilisée en France au XVIIIe siècle pour évaluer la récolte annuelle moyenne. En 1910 elle fut proposée par Poincaré comme une alternative plus stable de $\bar{x}$. Cette méthode est appliquée par les arbitres en gymnastique pour déterminer la note moyenne d'un concurrent. Mais plus tard, dans les années 30 à 40, les

brillants travaux de R. Fisher et les notions fondamentales d'exhaustivité (d'une statistique) et d'efficacité (d'une estimation) pour une classe bien définie d'hypothèses concurrentes occultèrent la stabilité des estimations. J. Tukey [136] fut le premier à attirer l'attention des statisticiens sur la stabilité des estimations.

Citons l'exemple d'une statistique stable, commode qui connut bien des déboires. Largement répandue au début du siècle, elle fut jugée pas assez efficace, disparut pratiquement des ouvrages de statistique mathématique dans les années 60 et ce n'est que dernièrement qu'elle a été remise à l'honneur. Nous avons dit l'estimation du paramètre d'échelle appelée erreur absolue moyenne $d_n = \sum_{i=1}^{n} | x_i - \bar{x} |/n$. Depuis les temps de Gauss, cette estimation a rivalisé avec l'estimation quadratique moyenne $s_n = \sqrt{\Sigma (x_i - \bar{x})^2/(n-1)}$. En général, d et s sont des paramètres différents. Pour la loi normale, le rapport de leurs valeurs limites est égal à : $\frac{d_\infty}{s_\infty} = \sqrt{2/\pi}$.

Pour comparer d_n et s_n, on se sert de la caractéristique asymptotique suivante qui exprime l'efficacité asymptotique de d_n par rapport à s_n comme une estimation du paramètre d'échelle :

$$e = \lim_{n \to \infty} \frac{\mathrm{V}(s_n/s_\infty)}{\mathrm{V}(d_n/d_\infty)},$$

et on effectue la comparaison sur la loi normale « polluée » de Tukey (cf. n° 6.1.11). Le tableau 10.3 représente les variations de e en fonction de ε. Sur ce tableau on voit que, dans le cas normal ($\varepsilon = 0$),

Tableau 10.3

ε	0	0,001	0,002	0,005	0,01	0,02	0,05	0,10	0,15	0,25
$e(\varepsilon)$	0,876	0,948	1,016	1,198	1,439	1,752	2,035	1,903	1,689	1,371

s est de 12 % plus efficace que d, mais déjà pour $\varepsilon = 0{,}002$, c'est d qui est plus efficace. L'efficacité de d continue de croître rapidement et, pour $\varepsilon = 0{,}05$, est de deux fois supérieure à celle de s. Comme d est plus efficace que s pour pratiquement toutes les valeurs de ε, il s'ensuit qu'il est préférable d'utiliser d pour les échantillons de petite et moyenne taille. On se servira de s uniquement pour les échantillons de très grande taille lorsque la répartition est très proche (!) de la normale ($\varepsilon < 0{,}002$).

Les estimations graphiques décrites au n° 10.4.3 sont manifestement stables pour les écarts de la répartition par rapport à la norma-

le, mais elles ne sont pas assez efficaces pour une loi normale. C'est qu'on voudrait tirer le plus grand profit de l'information contenue dans la partie centrale de la répartition. On peut le faire, par exemple, en utilisant des fonctions de poids spécialement choisies pour déterminer les paramètres. Dans le cas multidimensionnel, l'estimation M du vecteur de localisation et l'estimation $\mathbf{B}$ de la matrice des covariances peuvent être déterminées [33, 91] par la résolution itérative des équations

$$\sum_i (x_i - M)\, w_1(t_i) = 0\,; \tag{10.13}$$

$$\sum_i [(X_i - M)(X_i - M)' - \beta \mathbf{B}]\, w_2(t_i) = 0, \tag{10.14}$$

où $w_1(t)$, $w_2(t)$ sont des fonctions de poids données, $t_i = (X_i - M)' \mathbf{B}^{-1} (X_i - M)$, et β une constante choisie de manière à compenser dans le cas normal le biais de l'estimation de la matrice des covariances, causé par la pondération. Pour $w_1 = w_2 \equiv 1$ et $\beta = 1$, les formules (10.13) et (10.14) se ramènent aux habituelles estimations (10.11) et (10.12) de la loi normale.

10.4.5. Estimation de la position du centre des lois symétriques. Les anciennes et naïves hypothèses relatives à l'estimation du paramètre de localisation furent assises sur des bases nouvelles après le fameux travail de Tukey [136]. Décrivons les plus intéressantes d'entre elles. Dans la suite, $x_1, \ldots, x_n$ est un échantillon indépendant de fonction de répartition symétrique $P(x - \mu)$ et $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$ la série variationnelle correspondante.

La *moyenne tronquée d'ordre* α $(0 \leqslant \alpha < 0{,}5)$ pour $x_1, \ldots, x_n$ est définie par la formule

$$t(\alpha) = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)}, \tag{10.15}$$

où m est le plus grand entier $\leqslant \alpha n$. Sous certaines conditions imposées à la fonction de répartition $P(x)$, l'estimation est asymptotiquement normale et possède une variance asymptotique $n^{-1}\sigma_t^2(\alpha)$, où

$$\sigma_t^2(\alpha) = \frac{1}{(1-2\alpha)^2} \left(\int_{z_\alpha}^{z_{1-\alpha}} z^2 f(z)\, dz + 2\alpha z_\alpha^2 \right), \tag{10.16}$$

où $f(z)$ est la densité de $P(z)$ et $z_\alpha = P^{-1}(\alpha)$.

La *moyenne winsorienne d'ordre* α $(0 \leqslant \alpha < 0{,}5)$ pour $x_1, \ldots, x_n$ est donnée par la formule

$$w(\alpha) = \frac{1}{n} \left(\sum_{i=m+2}^{n-m-1} x_{(i)} + m\,(x_{(m+1)} + x_{(n-m)}) \right). \tag{10.17}$$

Cette estimation est aussi, sous certaines conditions portant sur la fonction de répartition $P(x)$, asymptotiquement normale de variance asymptotique $n^{-1}\sigma_w^2(\alpha)$, où

$$\sigma_w^2(\alpha) = \int_{z_\alpha}^{z_{1-\alpha}} z^2 f(z)\,dz + 2\alpha\left(z_{1-\alpha} + \frac{\alpha}{f(z_\alpha)}\right)^2. \qquad (10.18)$$

Les estimations $t(\alpha)$ et $w(\alpha)$ sont destinées à combattre les observations extrêmes qui sont considérées comme des erreurs grossières. Ces estimations remplissent bien leur contrat si les erreurs grossières sont aussi fréquentes à gauche et à droite de la série variationnelle. Mais si la répartition est asymétrique, il vaut mieux utiliser les estimations du n° 10.4.6. L'étude des propriétés des estimations du paramètre de localisation des répartitions symétriques a fait l'objet de nombreux travaux [71, 72, 91, 124]. Si l'on considère un modèle de « pollution » de la loi normale Φ de la forme

$$F(t-\mu) = (1-\varepsilon)\,\Phi(t-\mu) + \varepsilon H(t-\mu),$$

où $H(u)$ est la fonction de répartition d'une pollution arbitraire symétrique par rapport à $\mu = 0$, alors le minimum de la variance asymptotique maximale (sur H) de l'estimation de μ est atteint sur $t(\alpha)$, où le niveau de troncature α est choisi de telle sorte que $\alpha = \Phi(-k)$, où k est la solution de l'équation [73]:

$$\frac{\varepsilon k}{1-\varepsilon} = \frac{2}{\sqrt{2\pi}}\left(e^{-k^2/2} - k\int_k^\infty e^{-x^2/2}\,dx\right).$$

Ce résultat est intéressant sur le plan théorique; dans la pratique, ε est généralement inconnu et la répartition $H(x)$, rarement strictement symétrique.

La proposition de Humpel [71] d'estimer simultanément les paramètres de localisation m et d'échelle s moyennant la résolution de l'équation

$$\sum_i \Psi[(x_i - m)/s] = 0, \qquad (10.19)$$

où

$$\Psi(u) = \operatorname{sgn}(u)\begin{cases} |u|, & 0 \leqslant |u| \leqslant 1{,}7; \\ 1{,}7, & 1{,}7 \leqslant |u| < 3{,}4; \\ \dfrac{(8{,}5 - |u|)}{3}, & 3{,}4 \leqslant |u| < 8{,}5; \\ 0, & |u| \geqslant 8{,}5, \end{cases}$$

où s est la médiane des écarts absolus de x_i par rapport à m, va dans le sens des estimations exhibées au n° 10.4.6. Cette équation se ré-

sout par une méthode itérative. Pour approximation initiale de m, on prend la médiane x_i, et pour s, la médiane de l'écart absolu par rapport à m_0.

10.4.6. Paramétrisation à l'aide des estimations exponentiellement pondérées (EEP). Si dans les formules (10.13) et (10.14), on pose $w_1(t) = w_2(t) = \exp\{-\lambda t/2\}$, $\beta = (1+\lambda)^{-1}$, où λ est un petit paramètre, on obtient une classe à un paramètre d'estimations, proposée en 1970 par L. Méchalkine. Ces estimations sont commodes, stables pour les pollutions non symétriques, admettent une interprétation probabiliste simple et se généralisent au cas multidimensionnel. Ces dernières années, elles ont fait l'objet d'intenses études [33, 34, 96, 97] et on a élaboré pour elles une théorie asymptotique pour les répartitions normales, dans le cas multidimensionnel, et pour les répartitions non normales, dans le cas scalaire. L'interprétation géométrique suggestive des EEP et la simplicité des procédures de calcul nécessaires à leur mise en œuvre plaident pour leur généralisation aux problèmes de géométrie multidimensionnelle [83] et de régression [99].

Commençons l'exposé par l'interprétation probabiliste des EEP. Par une série de définitions, nous allons indiquer pour chaque répartition multidimensionnelle la loi normale la plus proche d'elle, et les paramètres de cette loi seront pris pour paramètres de la répartition initiale.

Soient $X \in R^p$, A un ensemble convexe de R^p et

$$\rho(F, G) = \sup_A \left| \int_A (\mathrm{d}F(X) - \mathrm{d}G(X)) \right|$$

la distance entre les répartitions F et G.

Définition 1. Etant donné une fonction de poids $w(X)$, on appellera *vecteur moyen w-pondéré* le vecteur

$$V = V_w(F) = \int X w(X)\, \mathrm{d}F(X)/e,$$

et matrice des covariances w-pondérée, la matrice

$$\Lambda = \Lambda_w(F) = \int (X - V)(X - V)' w(X)\, \mathrm{d}F(X)/e,$$

où $e = e_w(F) = \int w(X) \mathrm{d}F(X)$.

Définition 2. Les répartitions possédant les mêmes matrices des covariances w-pondérées seront dites *w-semblables*.

La notion de w-similitude permet de relier une loi arbitraire F à la loi normale N qui lui est w-semblable et à utiliser les moments d'ordre un et d'ordre deux de N pour décrire F. Le choix de la fonction de poids dépend essentiellement du problème posé. Si l'on se propose de décrire la partie centrale de la répartition, il est naturel

de relier la fonction de poids à la densité d'une loi normale w-semblable.

Définition 3. Soit $\varphi = \varphi(X; M, \Sigma)$ la densité en un point X d'une loi normale N de vecteur des moyennes M et de matrice des covariances Σ. On dira que la loi N est (λ, C)-*liée* (ou simplement λ-liée) à F si elle est φ^λ-semblable à F et $\rho(F, N) < C$. Cette dernière condition a été introduite pour assurer l'unicité de la loi normale λ-liée à F pour de petits C, car dans le cas général il peut exister plusieurs lois normales φ^λ-semblable à F.

Définition 4. Soit N une loi normale λ-liée à F. Le vecteur des moyennes et la matrice des covariances de N seront appelés respectivement *λ-vecteur moyen* et *λ-matrice des covariances* de F.

De ces définitions il s'ensuit, en particulier, que les λ-moments de toute loi normale sont confondus avec ses moments ordinaires.

Soient $\mathbf{N}(p)$ l'ensemble de toutes les lois normales p-dimensionnelles non dégénérées, c'est-à-dire non concentrées dans un sous-espace de dimension $< p$, et $\mathbf{M}(p, \varepsilon)$ l'ensemble de toutes les répartitions F, telles que

$$\rho(F, \mathbf{N}(p)) = \inf_{N \in \mathbf{N}(p)} \rho(F, N) < \varepsilon.$$

En se référant aux travaux [33, 97], on démontre que, pour tout $\lambda > 0$, il existe des $C = C(\lambda, p)$ et $\varepsilon = \varepsilon(p, \lambda, C) > 0$, tels que pour tout $F \in \mathbf{M}(p, \varepsilon)$:

1) il existe une loi normale et une seule λ-liée à F;

2) le λ-vecteur moyen (M_F) et la λ-matrice des covariances (Σ_F) sont des fonctions continues de F (au sens de la ρ-distance);

3) le λ-vecteur moyen et la λ-matrice des covariances de F varient comme les moments respectifs de la loi normale par toute transformation linéaire des variables;

4) les λ-moments satisfont le système d'équations

$$M = \int X\varphi^\lambda(X, M, \Sigma)\,dF(X) \Big/ \int \varphi^\lambda(X, M, \Sigma)\,dF(X); \quad (10.20)$$

$$\Sigma = (1+\lambda) \int (X-M)(X-M)'\varphi^\lambda \times$$
$$\times (X, M, \Sigma)\,dF(X) \Big/ \int \varphi^\lambda(X, M, \Sigma)\,dF(X), \quad (10.21)$$

où $\varphi(x, M, \Sigma) = \exp\{-(x-M)'\Sigma^{-1}(x-M)/2\}$. Si la fonction $F(x)$ est connue et si l'approximation initiale est convenablement choisie, on peut résoudre le système (10.20), (10.21) itérativement;

5) si l'on estime les paramètres sur le vu d'un échantillon indépendant de taille n, les intégrales du système (10.20), (10.21) doivent être remplacées par les sommes correspondantes étendues à toutes les

observations :

$$\hat{M} = \sum_i X_i \varphi^\lambda (X_i, \hat{M}, \mathbf{B}) / \sum_i \varphi^\lambda (X_i, \hat{M}, \mathbf{B}); \qquad (10.20')$$

$$\mathbf{B} = (1+\lambda) \sum_i (X_i - \hat{M})(X_i - M)' \varphi^\lambda (X_i, \hat{M}, \mathbf{B}) / \sum_i \varphi^\lambda (X_i, \hat{M}, \mathbf{B}). \qquad (10.21')$$

Pour le cas normal, A. Chouryguine [33] préconise, pour compenser la finitude de l'échantillon, d'introduire au second membre de (10.21′) le facteur $(1 + \tau/n)^{-1}$, où $\tau = -(1+\lambda)^{p+4}(1 + 2\lambda)^{-p/2-1} + \lambda(1+\lambda)^{p+1}(1+2\lambda)^{-p/2-3}[(5 + 14\lambda + 15\lambda^2 + 8{,}5\lambda^3 + 4\lambda^4) + (p-1)(1 - \lambda - 4\lambda^2 - 3{,}5\lambda^3) - \lambda(p-1)(p-2) \times (0{,}5 + \lambda + 0{,}5\lambda^2)]$;

6) les estimations $\hat{M}$ et $\mathbf{B}$ sont convergentes en probabilité et asymptotiquement normales. En dimension un, leur matrice des covariances asymptotique est de la forme $\mathbf{C}n^{-1}$, où $\mathbf{C}$ peut être exprimé en fonction des quatre premiers λ- et 2λ-moments. Si $a_k(\lambda) = \int (X - M)^{(k)} \varphi^\lambda (X, M, \Sigma)\, dF(X)$, alors $\mathbf{C} = \mathbf{K}^{-1}\mathbf{H}\mathbf{K}^{-1}$, où $\mathbf{K}$ et $\mathbf{H}$ sont des matrices carrées symétriques d'ordre 2 ;

$$k_{11} = 1, \qquad k_{12} = -\frac{\lambda(1+\lambda)}{2\Sigma^2} \cdot \frac{a_3(\lambda)}{a_0(\lambda)},$$

$$k_{22} = \frac{2+3\lambda}{4\Sigma(1+\lambda)} - \frac{\lambda(1+\lambda)}{4\Sigma^3} \cdot \frac{a_4(\lambda)}{a_0(\lambda)}, \qquad h_{11} = (1+\lambda)^2 \cdot \frac{a_2(2\lambda)}{a_0^2(\lambda)},$$

$$h_{12} = \frac{1+\lambda}{2} \cdot \left(\frac{a_3(2\lambda)}{4\Sigma} - \frac{a_1(2\lambda)}{a_0^2(\lambda)} \right),$$

$$h_{22} = \left(\frac{a_4(2\lambda)}{a_2^2(\lambda)} + \frac{a_0(2\lambda)}{a_0^2(\lambda)} - 2\frac{a_2(2\lambda)}{a_2(\lambda)\, a_0(\lambda)} \right) \Big/ 4.$$

Dans les calculs pratiques, les quantités $a_k(\lambda)$ peuvent être remplacées par leurs estimations sur le vu des données empiriques ;

7) dans le cas normal multidimensionnel, les estimations $\hat{M}$ et $\mathbf{B}$ sont asymptotiquement indépendantes, la matrice des covariances asymptotique de $\hat{M}$ est de la forme : $n^{-1}(1+\lambda)^{p+2}(1+2\lambda)^{-p/2-1}\Sigma$.

Les propriétés empiriques des λ-moments sont illustrées pour la dimension un sur le tableau 10.4 [98] qui représente les estimations de la λ-moyenne et de la λ-variance calculées sur 100 échantillons de taille 100 *). On voit sur ce tableau que l'introduction d'une fonction de poids dans le cas normal altère un peu les propriétés des estimations, mais par contre, dans le cas d'une répartition « polluée » non seulement elle réduit le biais des estimations, mais elle améliore leurs propriétés empiriques ($\lambda = 0{,}5$). Dans la pratique, il est re-

*) Les chiffres précédés du signe ± expriment la valeur empirique de l'écart-type de l'estimation correspondante.

Tableau 10.4

Loi	$N(0,1)$		$0,9\,N(0,1)+0,1\,N(1,3)^2$	
paramètre / λ	λ-moyenne	λ-variance	λ-moyenne	λ-variance
0	0,000±0,010	1,011±0,013	0,100±0,012	1,874±0,046
0,5	0,006±0,011	1,030±0,016	0,020±0,011	1,178±0,020
1	0,010±0,013	1,026±0,023	0,020±0,013	1,118±0,025
1,5	0,014±0,015	1,015±0,030	0,023±0,015	1,083±0,031

commandé de choisir les valeurs de λ en fonction de la taille de l'échantillon et de la dimension p de l'espace des échantillons de telle sorte que, d'une part, la pondération « amortisse » les grands écarts et, de l'autre, que la perte d'efficacité due à cette pondération ne soit pas trop grande.

Les estimations pondérées avec des poids arbitraires sont traitées dans [91].

10.5. Visualisation des données multidimensionnelles

10.5.1. Position du problème. Comme indiqué au § 1.1, les données recueillies peuvent souvent être traitées comme des vecteurs (des points) d'un espace à plusieurs dimensions. Si le modèle mathématique de la situation étudiée est connu, on peut s'imaginer avec plus ou moins de précision comment les points-observations seront disposés dans cet espace. Mais, en règle générale, avant d'étudier la disposition des points, on formule les hypothèses et on construit les modèles. Les méthodes mathématiques sollicitées pour étudier la position des points reposent sur la notion de distance entre eux. Le travail [9] nous fournit plusieurs exemples de distances. Les méthodes de visualisation des données sont toutes basées sur les faits intuitifs suivants:

1) « la dimension réelle de la situation étudiée » est très inférieure à l'espace des échantillons;

2) la position relative des points peut être étudiée par un assez vaste spectre de distances. L'expression « dimension réelle de la situation » appelle quelques précisions. A défaut d'une définition exacte, on se propose de l'expliquer par un exemple.

Exemple 10.3. On fait souvent appel à l'analyse discriminante linéaire fishérienne [9, 16, 103, 135] pour distinguer deux groupes d'objets dans un espace d'échantillons à p dimensions et classer toute nouvelle observation dans l'un de ces groupes. Dans ce cas, on suppose que les répartitions p-dimensionnelles des objets de chaque

groupe sont des échantillons de deux répartitions normales p-dimensionnelles de matrice des covariances commune et de vecteurs des moyennes différents. L'ensemble de toutes les répartitions des données initiales se réduit donc à deux points de l'espace des observations : à la moyenne des groupes respectifs et à une seule matrice des covariances, c'est-à-dire, en fait, à un problème en dimension un. Cette circonstance facilite considérablement la manipulation des données et permet d'estimer sans peine la probabilité qu'une nouvelle observation appartienne à tel ou tel groupe.

Les méthodes de visualisation des données sont basées sur le passage des variables initiales $x^{(1)}, \ldots, x^{(p)}$ à un nouveau système de variables qui sont des fonctions des variables primitives. Il est souhaitable qu'un nombre petit q (pour la visualisation q peut être égal soit à 1, soit à 2, soit à 3) de nouvelles variables conservent les traits les plus caractéristiques de la structure de la matrice des données, par exemple les « clusters », les « chaînes », les objets éloignés de la principale population et autres formations.

Dans le cas général, le problème du passage (avec une perte minimale d'information) des variables initiales $x^{(1)}, \ldots, x^{(p)}$ aux nouvelles variables $z^{(1)}, \ldots, z^{(q)}$ $(q \ll p)$ est présenté au § 1.2 comme un problème d'extremum, dans lequel le choix des nouvelles variables $Z(X) = (z^{(1)}(X), \ldots, z^{(q)}(X))'$ passe par la maximisation d'une mesure d'informativité $I_q(Z(X))$ définie de façon exogène.

Le choix de la fonctionnelle $I_q(Z)$ dépend des objectifs de la visualisation et de l'information *a priori* dont on dispose sur la structure de la population d'objets étudiée. Si cette information fait défaut, on utilise des critères dits d'*autoinformativité*, qui visent à conserver au maximum l'information contenue dans la matrice des données initiales. La *méthode des composantes principales* et ses généralisations non linéaires sont basées sur ces critères. L'information *a priori* concerne souvent la non-homogénéité de la population des objets, c'est-à-dire l'appartenance des objets à tel ou tel groupe, par exemple au groupe des personnes malades ou au groupe des personnes saines, ce qui implique de faire appel à des critères de l'analyse discriminante visant à préserver cette information.

Comme indiqué au § 1.1, les données initiales peuvent être définies aussi sous la forme d'une matrice des proximités (ou d'une matrice des distances) entre les objets. La visualisation des données est dans ce cas étroitement liée aux méthodes de l'analyse des proximités [75] et implique la construction d'une matrice des données (la configuration des points) éventuellement de moindre dimension qui « expliquerait » la matrice des distances.

On étudie plus bas deux méthodes de visualisation des données multidimensionnelles :

1) la méthode des composantes principales qui ramène le problè-

me à la projection des points sur les plans construits sur les premiers vecteurs propres de leur matrice des covariances commune;

2) la représentation de la matrice des distances (cf. § 1.1) par un nuage de points dans un espace de petite dimension.

10.5.2. Composantes principales. On appelle *composantes principales* $z^{(1)}, \ldots, z^{(p)}$ de la variable $X = (x^{(1)}, \ldots, x^{(p)})'$ le système de combinaisons linéaires orthonormales des variables initiales

$$\begin{cases} z^{(j)}(X) = u_{1j}(x^{(1)} - m^{(1)}) + \ldots + u_{pj}(x^{(p)} - m^{(p)}); \\ \sum_{i=1}^{p} u_{ij}^2 = 1 \quad (j = 1, \ldots, p); \\ \sum_{i=1}^{p} u_{ij}u_{ik} = 0 \quad (j, k = 1, \ldots, p, j \neq k), \end{cases} \tag{10.22}$$

où $m^{(j)}$ est l'espérance mathématique de $x^{(j)}$. Dans la suite, si le contraire n'est pas spécifié, on admettra que les données sont centrées, c'est-à-dire que $m^{(j)} = 0$. Les combinaisons linéaires sont choisies de la manière suivante. La première composante principale possède la plus grande variance parmi les combinaisons linéaires (10.22). Géométriquement cela signifie que la première composante principale est orientée dans le sens du plus grand allongement de l'ellipsoïde de dispersion de la population étudiée. La deuxième composante principale admet la plus grande variance parmi les combinaisons linéaires (10.22) non corrélées à la première composante. C'est la projection sur la direction du plus grand allongement des observations de l'hyperplan perpendiculaire à la première composante principale. La troisième composante principale possède la plus grande dispersion parmi les combinaisons (10.22) non corrélées avec les deux premières composantes principales, et ainsi de suite. De façon plus formelle, cela signifie que l'ensemble des q premières composantes principales est optimal pour le critère

$$I_q(Z(X)) = \frac{\sum_{j=1}^{q} \mathrm{V}z^{(j)}}{\sum_{j=1}^{p} \mathrm{V}x^{(j)}}, \tag{10.23}$$

où $\mathrm{V}z^{(j)}$ est la variance de la j-ième composante principale; $\mathrm{V}x^{(i)}$ la variance de la variable $x^{(i)}$.

Le calcul des coefficients des composantes principales est basé sur le fait que les vecteurs $U_1, \ldots, U_q$ sont les vecteurs propres (cf. [16]) de la matrice des covariances de la population étudiée, c'est-à-dire satisfont le système d'équations

$$\Sigma U = \lambda U.$$

Les valeurs propres correspondantes sont égales aux variances des composantes principales. Si maintenant l'on range les vecteurs propres dans l'ordre de grandeur décroissante des valeurs propres $\lambda_1 \geqslant \ldots \geqslant \lambda_p$, à la première composante principale sera associé le vecteur propre correspondant à la plus grande valeur propre λ_1 et $Vz^{(1)} = \lambda_1$; à la deuxième composante principale sera associé le vecteur propre correspondant à la valeur propre suivante λ_2 et $Vz^{(2)} = \lambda_2$, et ainsi de suite.

Il existe en tout p vecteurs propres, donc, p composantes principales. La matrice des covariances des composantes principales $z^{(1)}, \ldots, z^{(p)}$ sera de la forme

$$\Sigma_Z = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix}.$$

Vu que la transformation U qui permet de passer des variables initiales aux composantes principales est orthogonale, on a les relations suivantes

$$\operatorname{Tr} \Sigma = \sum_{i=1}^{p} Vx^{(i)} = \sum_{i=1}^{p} Vz^{(i)} = \sum_{i=1}^{p} \lambda_i;$$

$$\det \Sigma_Z = \det \Sigma,$$

c'est-à-dire que la variance généralisée det Σ_Z et la somme des variances des composantes principales sont respectivement égales à la variance généralisée det Σ et à la somme des variances des variables initiales. Donc, pour les q premières composantes principales, le critère (10.23) est égal à

$$I_q(Z) = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, \tag{10.23'}$$

c'est-à-dire au pourcentage de variance totale « explicable » par les q premières composantes principales. Plus la valeur de ce critère sera proche de l'unité et moins on a de chance d'altérer la position relative des points empiriques en les projetant sur l'hyperplan construit sur les q premières composantes principales.

Remarque 1. En pratique, la valeur théorique de Σ est la plupart du temps inconnue, aussi considère-t-on à sa place la matrice des covariances empirique $\hat{\Sigma} = XX'$ et parle-t-on des composantes principales de l'échantillon. Si aucune confusion n'est à craindre, on omettra le terme « empirique ». Signalons encore que le terme « com-

posante principale » désigne aussi bien le scalaire $z^{(j)}$ que le vecteur propre U_j correspondant.

Remarque 2. La méthode des composantes principales est la plus féconde lorsque toutes les composantes du vecteur étudié admettent la même signification physique et sont respectivement exprimées dans les mêmes unités. Si les variables sont mesurées dans des unités différentes, le résultat risque de dépendre essentiellement du choix de l'échelle de mesure. Pour remédier à cette situation, on passe souvent à des variables sans dimension en normant $x^{(i)}$, par exemple par la transformation $x^{*(i)} = x^{(i)}/\hat{\sigma}_{ii}$, où σ_{ii} est la variance de $x^{(i)}$.

10.5.3. Propriétés de la plus petite déformation de la structure géométrique des composantes principales. Exhibons deux assertions concernant l'invariance de la structure géométrique locale de l'ensemble des points initiaux par leur projection sur les q premières composantes $U_1, \ldots, U_q$.

1. Soit L un hyperplan arbitraire à q dimensions ($q < p$) passant par l'origine des coordonnées, et soit $Z_1, \ldots, Z_n$ les projections de $X_1, \ldots, X_n$ sur L. Alors la quantité $\Delta(L) = \sum_{i=1}^{n} (X_i - Z_i)'(X_i - Z_i)$, qui est égale à la somme des carrés des écarts de X_i par rapport à L, atteint son minimum lorsque L est confondu avec l'hyperplan C construit sur les q premières composantes principales. Ceci étant

$$\Delta_q(C) = \min_L I(L) = n(\hat{\lambda}_{q+1} + \ldots + \hat{\lambda}_p). \tag{10.24}$$

2. Considérons par ailleurs la matrice $\mathbf{H}$ d'ordre n et d'éléments $h_{ij} = \sum_{k=1}^{p} x_i^{(k)} x_j^{(k)}$ égaux au produit scalaire des vecteurs X_i et X_j, et soit $\mathbf{H}(L)$ la matrice analogue construite sur les vecteurs $Z_1, \ldots, Z_q$. La signification géométrique de ces matrices est évidente: h_{jj} est le carré de la longueur du vecteur X_j, h_{ij} est proportionnel au cosinus de l'angle de X_j et X_i. On a

$$\|\mathbf{H} - \mathbf{H}(C)\| = \min_L \|\mathbf{H} - \mathbf{H}(L)\| = n^2(\hat{\lambda}_{q+1}^2 + \ldots + \hat{\lambda}_p^2) \tag{10.25}$$

où $\|\mathbf{H}\| = \sqrt{\sum_{i,j} h_{ij}^2}$, autrement dit, l'hyperplan construit sur les q premières composantes principales déforme le moins la longueur et les angles des projections.

De (10.24) et (10.25) il s'ensuit que, pour mesure de l'invariance des propriétés géométriques des objets par leur projection sur L,

on peut adopter soit la quantité

$$I_q(Z) = \frac{\sum_{i=1}^{p-q} \hat{\lambda}_{q+i}}{\sum_{i=1}^{p} \hat{\lambda}_i},$$

soit la quantité

$$\tilde{I}_q(Z) = \frac{\sum_{i=1}^{p-q} \hat{\lambda}_{q+i}^2}{\sum_{i=1}^{p} \hat{\lambda}_i^2}.$$

Les programmes réalisant la méthode des composantes principales figurent dans pratiquement tous les paquets de programmes statistiques. Les principaux défauts de la méthode des composantes principales sont liés au fait que, *primo*, l'estimation $\hat{\Sigma}$ peut être faussée par des « anomalies » passées inaperçues dans les données, et, *secundo*, cette méthode vise essentiellement à dégager les relations linéaires.

On vient à bout du premier des problèmes indiqués en passant aux estimations stables, par exemple aux estimations pondérées (cf. n° 10.4.6 et [83]), ou encore en supprimant les anomalies à l'aide, par exemple, des diagrammes de dispersion (scattergrammes). On peut aussi procéder à l'estimation non pas sur l'échantillon tout entier mais sur une de ses parties. En médecine, par exemple, on peut utiliser les seules données des personnes saines. De façon analogue, pour améliorer la suggestivité d'un diagramme de dispersion mettant en jeu un grand nombre d'objets, l'analyste aura intérêt à projeter non pas toutes les observations mais seulement celles qui l'intéressent en premier chef.

Pour remédier au deuxième inconvénient, on peut appliquer une méthode non linéaire d'application des données dans un espace de moindre dimension.

10.5.4. Applications non linéaires dans un espace de moindre dimension. Soient $X_1, \ldots, X_n$ des observations dans un espace à p dimensions, $Z_1, \ldots, Z_n$ leurs images par une application bijective non linéaire dans un espace à q ($q < p$) dimensions. Pour mesure de la qualité de cette application, prenons une fonctionnelle reflétant l'invariance de la structure géométrique locale de la matrice des données $\mathbf{X}$. On trouvera plusieurs types de telles fonctionnelles dans [9, 62, 130]. On se bornera à étudier la fonctionnelle

$$I(\mathbf{Z}, a) = \sum_{i<j}^{n} (d_{ij} - d_{ij}^*)^2 d_{ij}^a / \sum_{i<j}^{n} d_{ij}, \qquad (10.26)$$

où d_{ij} est la distance sur l'espace des observations X et d^*_{ij} la distance sur l'espace d'arrivée Z.

Pour distance on peut prendre la distance euclidienne pondérée

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^{p} w_k (x_i^{(k)} - x_j^{(k)})^2\right)^{1/2}$$

(la distance euclidienne ordinaire s'en déduit comme un cas particulier pour $w_1 = w_2 = \ldots = 1$); la distance de Mahalanobis est invariante par les applications linéaires de l'espace de départ,

$$d_{ij} = (X_i - X_j)' \hat{\Sigma}^{-1} (X_i - X_j),$$

et la distance de Kolmogorov,

$$d_{ij} = \left(\sum_{k=1}^{p} w_k \mid x_i^{(k)} - x_j^{(k)} \mid^p\right)^{1/p}, \qquad p > 0.$$

L'espace d'arrivée peut également être muni de l'une des métriques précédentes.

Dans la suite, les espaces de départ et d'arrivée seront munis de la métrique euclidienne. On passe aisément de la distance euclidienne pondérée et de la distance de Mahalanobis à la distance euclidienne ordinaire par une transformation linéaire des données.

Supposons donc que l'espace d'arrivée est muni de la distance

$$d^*_{ij} = \left(\sum_{i=1}^{q} (z_i^{(k)} - z_j^{(k)})^2\right)^{1/2}.$$

En portant cette expression dans (10.26), on met la fonctionnelle $I(\mathbf{Z}, a)$ sous la forme d'une fonction de $q \times n$ coordonnées $\mathbf{Z}_i^{(k)}$ ($i = 1, \ldots, n$, $k = 1, \ldots, q$), et la minimisation de la fonctionnelle $I(\mathbf{Z}, a)$ se ramène à celle d'une fonction de $q \times n$ variables.

Avant de passer à la description de la procédure de minimisation, considérons les propriétés du critère (10.26) en fonction des valeurs du paramètre a. Si $a > 0$, la déformation de la distance (c'est-à-dire la différence $(d_{ij} - d^*_{ij})^2$) se répercute d'autant plus fortement sur la valeur du critère que la distance entre les points de l'espace de départ est élevée. C'est pourquoi, en se servant d'un critère dont le paramètre $a > 0$, on observera que plus la distance entre les points de l'espace de départ est élevée, et moins la distance de leurs images sera déformée. Pour les mêmes raisons, si $a < 0$, ce sont les images des petites distances qui seront le moins déformées.

L'approche la plus souple est celle qui utilise un paramètre a qui varie en fonction de la différence $d_{ij} - d^*_{ij}$. Plus exactement,

$$a = \begin{cases} -a_1 & \text{si} \quad d_{ij} \leqslant d^*_{ij}; \\ a_2 & \text{si} \quad d_{ij} > d^*_{ij}. \end{cases} \tag{10.27}$$

Traitons le cas où $a_1 > 0$ et $a_2 > 0$. Supposons que $d_{ij} < d^*_{ij}$ et que d_{ij} est très grand. Alors, le poids de la déformation $(d_{ij} - d^*_{ij})^2$ sera petit $(1/d^{a_1}_{ij})$. D'autre part, si $d_{ij} > d^*_{ij}$ et d_{ij} est encore très grand, alors le poids de la déformation $(d_{ij} - d^*_{ij})^2$ sera élevé $(d^{a_1}_{ij})$. Donc, en minimisant $I(Z, a)$, on constatera que la déformation des grandes distances a tendance à croître, et celle des petites, à décroître. Il est normal qu'avec des déformations de cette nature l'image de la configuration de départ soit meilleure. En tout cas, si l'on a affaire à des clusters, ce type de déformations peut être plus utile dans la mesure où les distances entre les points d'un même cluster auront tendance à être petites, et les distances entre les points de clusters différents, à être grandes, donc, les déformations de ce type accentueront le « contraste » de la visualisation.

Comme déjà signalé, la détermination de la configuration, pour laquelle la fonctionnelle de qualité I (Z, a) prend sa valeur minimale, se ramène à un problème de minimisation d'une fonction de $q \times n$ variables. Vu le nombre élevé de variables, les méthodes de résolution les plus performantes seront des modifications de la méthode du gradient.

Considérons une procédure de résolution réalisée dans le programme de réduction de la dimension dans le paquet des PPSA [108]. Le minimum de la fonction I (Z, a) dans un espace à $q \times n$ dimensions est déterminé par la procédure itérative:

$$Z^{(k)}_{i(t+1)} = Z^{(k)}_{i(t)} - \frac{1}{2\sum\limits_{i>j} d^a_{ij}} \left[\frac{\partial I}{\partial z^{(k)}_i}\right]_{(t)} \quad (i = 1, \ldots, n; \quad k = 1, \ldots, q),$$

où t est le numéro du pas de l'itération, $Z^{(k)}_i$ la k-ième coordonnée de l'image de X_i, et

$$\frac{\partial I}{\partial z^{(k)}_i} = 2 \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(1 - \frac{d_{ij}}{d^*_{ij}}\right) d^a_{ij}(z^{(k)}_i - z^{(k)}_j). \tag{10.28}$$

Donc, c'est une procédure de type gradient. Elle se caractérise par le choix d'un pas égal à $\frac{1}{2}\sum\limits_{i>j} d^a_{ij}$. On justifie la valeur de ce pas comme on a justifié celle du pas choisi pour le critère (10.26) (avec $a = 0$) cité dans le travail [75] (ce pas est visiblement égal à $1/2n$). On démontre que la procédure itérative (10.28) converge vers un point de minimum local de la fonctionnelle I (Z, a) quelles que soient les conditions initiales. Etant donné que la fonctionnelle I (Z, a) est susceptible d'avoir un très grand nombre de minimums locaux, pour tomber sur son minimum absolu, il faut exécuter de nombreux calculs qui occuperont assez longtemps l'ordinateur. Ceci donne un relief tout particulier au choix de l'approximation initiale. Dans le

paquet des PPSA [108], pour approximation initiale on a pris la projection des points de X sur les q premières composantes principales. Si $a = 0$, les composantes principales minimisent le critère I (Z, a) dans la classe des applications linéaires orthogonales. Même si ce choix des conditions initiales ne garantit pas l'atteinte du minimum absolu, il n'en contribue pas moins dans de nombreux cas à améliorer considérablement la configuration image par rapport à la méthode des composantes principales. Signalons en conclusion une importante propriété de l'application non linéaire: le centre de gravité de la configuration image est confondu avec celui de la configuration prise pour approximation initiale.

10.5.5. Analyse des proximités linéaire. Dans des numéros précédents nous avons étudié le problème de la visualisation des données sous forme d'une matrice. Le fichier est parfois représenté par une $n \times n$-matrice des distances **D** entre les objets. Plaçons-nous dans le cas où les éléments de la matrice **D** sont traités comme les distances entre les objets et essayons de construire la configuration de plus petite dimension susceptible d'engendrer (d'expliquer) la matrice **D**. Ce problème s'appelle *analyse des proximités linéaire.* Pour le résoudre, on se servira de nouveau d'une fonctionnelle (10.26). La procédure itérative de détermination du minimum est la même. En revanche, la donnée d'une configuration initiale heureuse se complique sensiblement. Voyons une méthode qui conduit à choisir une configuration initiale équivalente à celle obtenue par la méthode des composantes principales. Cette méthode est réalisée dans les PPSA.

Suivant [133] introduisons une matrice symétrique $\mathbf{H}^i$ d'ordre n d'éléments

$$h_{jk}^i = \frac{1}{2}(d_{ij}^2 + d_{ik}^2 - d_{jk}^2) \quad (j = 1, \ldots, n; \quad k = 1, \ldots, n). \tag{10.29}$$

La valeur de i est supposée fixée. Les éléments de la i-ième ligne et de la i-ième colonne sont supposés nuls. L'élément h_{jk}^i peut être interprété comme le produit scalaire des vecteurs d'origine, le point d'indice i, et d'extrémités, les points d'indices j et k. En effet, pour les points X_i, X_j, X_k, on a

$$d_{jk}^2 = d_{ij}^2 + d_{ik}^2 - 2d_{ij}\, d_{ik} \cos \theta_{jk}^i,$$

où θ_{jk}^i est l'angle des vecteurs $X_k - X_i$ et $X_j - X_i$. D'où

$$d_{ij}\, d_{ik} \cos \theta_{jk}^i = \frac{1}{2}(d_{ij}^2 + d_{ik}^2 - d_{jk}^2). \tag{10.30}$$

En comparant (10.29) et (10.30) on trouve que $h_{jk}^i = (X_j - X_i)' \times \times (X_k - X_i)$, $X_j - X_i$ et $X_k - X_i$ d'origine X_i et d'extrémités X_j et X_k. Il existe autant de matrices $\mathbf{H}^i$ que de points X_i, c'est-à-

dire n. On peut définir maintenant la matrice des produits scalaires des vecteurs issus du centre de l'ensemble des points $X_1, \ldots, X_n$, c'est-à-dire la matrice d'ordre n d'éléments

$$h_{jk} = (X_j - \overline{X})' (X_k - \overline{X}).$$

Les éléments de cette matrice se déduisent de ceux de n'importe quelle matrice $\mathbf{H}^i$ par la formule

$$h_{jk} = h^i_{jk} - \frac{1}{n} \sum_{j=1}^{n} h^i_{jk} - \frac{1}{n} \sum_{k=1}^{n} h^i_{jk} + \frac{1}{n^2} \sum_{j,k} h^i_{jk}. \qquad (10.31)$$

La matrice $\mathbf{H}^i$ est justement la matrice qui sera utilisée pour déterminer la configuration initiale. En admettant que les vecteurs sont centrés, on peut écrire $\mathbf{H} = \mathbf{X}'\mathbf{X}$. Donc la matrice $\mathbf{H}$ possède les mêmes rang et valeurs propres que la matrice des covariances $\hat{\Sigma}$ (qui est inconnue). Chaque vecteur propre de la matrice $\mathbf{H}$ représente la composante principale correspondante de $\mathbf{X}$, c'est-à-dire que la i-ième composante ($i = 1, \ldots, n$) du j-ième vecteur propre de $\mathbf{H}$ est le produit scalaire de X_i par le j-ième vecteur propre de la matrice $\hat{\Sigma}$. Si à des fins de visualisation on pose, par exemple, $q = 2$ et on prend les deux premiers vecteurs propres de la matrice $\mathbf{H}$ pour coordonnées de la configuration initiale, on se retrouve dans la situation décrite dans le numéro précédent.

Conclusions

1. Le succès d'une analyse, *a fortiori* si elle est volumineuse, passe par l'établissement d'un dossier d'étude décrivant avec soin l'organisation de la collecte des données, les variables enregistrées et le déroulement des opérations. L'état actuel du logiciel permet d'atteindre un niveau élevé d'automatisation et d'exécution de ce processus. L'analyse statistique doit être précédée d'une inspection minutieuse des données.

2. Les mesures peuvent être effectuées dans des échelles différentes : nominale, ordinale, d'intervalles et des rapports. Chaque type d'échelle implique sa propre technique de traitement statistique. Il existe des méthodes mathématiques spéciales axées sur l'utilisation de données multidimensionnelles exprimées dans des échelles différentes. Mais la situation la plus typique de l'analyse multidimensionnelle est celle où les échelles sont de même type. Pour se placer dans ce cadre, il faut soit passer à des variables prenant les seules valeurs 0 et 1, soit coder les variables nominales et ordinales.

3. La méthode la plus répandue d'estimation de la densité de probabilité d'une variable aléatoire est la construction de l'histogramme de cette variable ou du polygone de fréquence. Les données

groupées permettent d'estimer également les moments de la répartition, à condition d'introduire des corrections sur le groupement. L'assistance d'un ordinateur permet d'aborder les estimations non paramétriques de la densité, basées sur la moyennisation des « poids » de certaines observations dispersées autour de chaque observation par le biais d'une fonction de poids spécialement choisie.

4. On utilise souvent des transformations spéciales pour ramener la loi de probabilité des variables aléatoires étudiées à la loi normale. Dans ce cas, la densité de probabilité de la variable primitive est égale à celle de la loi normale de la variable image, multipliée par le jacobien de la transformation (cf. aussi (7.11)).

5. Pour représenter graphiquement la fonction de répartition, on se sert de papiers spéciaux, sur lesquels les fonctions de répartition de la forme $F((x - \mu)/\sigma)$, où F est connue, et μ et σ sont des paramètres inconnus, se représentent par des droites. Le plus courant de ces papiers est le papier gausso-arithmétique.

6. Pour représenter les répartitions multidimensionnelles de variables aléatoires mesurées dans des échelles nominale et ordinale, on fait largement appel aux tableaux de contingence dont l'élément $i, j, \ldots, k$ indique combien d'observations de l'échantillon possèdent la modalité i de la première variable, la modalité j de la seconde, . . ., la modalité k de la dernière. L'hypothèse la plus souvent testée dans les tableaux à deux dimensions est l'indépendance des lois de probabilité de la première et de la deuxième variable.

7. Quand on manipule des données multidimensionnelles, on a souvent besoin d'avoir une idée générale sur la disposition relative des points-observations dans l'espace des variables correspondant. Toutes les méthodes sont basées sur la notion de « distance » entre les points-observations. L'une des plus courantes est la méthode des composantes principales qui permet de projeter les points-observations sur la direction du plus grand allongement des données.

L'assistance des ordinateurs a permis de développer largement des méthodes directes de recherche, dans un espace de dimension peu élevé, d'une configuration de points, telle que la distance entre les points de cet espace corresponde le plus (le degré de cette correspondance est mesuré par une fonctionnelle spécialement choisie) à la distance entre les points-objets. Ces méthodes ont reçu le nom d'analyse des proximités. Elles sont encore peu élaborées sur le plan probabiliste et statistique.

8. Pour décrire les répartitions proches de la répartition normale, on se sert traditionnellement de la moyenne arithmétique et de l'écart-type. Mais cette approche n'est pas toujours heureuse, car les estimations indiquées perdent rapidement leurs bonnes propriétés au moindre écart par rapport à la répartition normale. Dans cette situation, les méthodes qui sont moins efficaces pour la répartition normale, mais plus stables (plus « robustes ») aux écarts par rapport

à la répartition normale prennent le pas sur les autres. Signalons à ce sujet les méthodes graphiques et l'estimation, oubliée depuis belle lurette, du paramètre d'échelle, qui est basée sur l'écart absolu moyen.

De nombreuses estimations stables des paramètres ont été proposées ces vingt dernières années. Les plus remarquables d'entre elles sont les méthodes qui utilisent, pour calculer les moments, la pondération des observations à l'aide de la densité d'une loi normale convenablement choisie.

CHAPITRE 11

ANALYSE PRÉLIMINAIRE DE LA NATURE DES DONNÉES

Au stade initial du traitement statistique des données initiales

$$X_1, X_2, \ldots, X_n \tag{11.1}$$

qui sont les valeurs observées de la variable aléatoire étudiée ξ, on doit donner aux questions suivantes des réponses qui conditionneront le choix des méthodes d'analyse les plus efficaces:

peut-on admettre que les données disponibles (11.1) sont les résultats d'observations indépendantes d'une variable aléatoire?

peut-on admettre que les échantillons initiaux sont extraits d'une même population générale?

est-il légitime de conjecturer que la loi de probabilité de la variable aléatoire ξ est symétrique par rapport au centre de groupement des valeurs de ξ?

quel modèle est le plus approprié pour décrire la loi de probabilité des données traitées, et, en particulier, comment vérifier l'accord du modèle choisi avec les observations disponibles?

comment traiter les données si certaines viennent à manquer (par exemple, certaines composantes d'une observation multidimensionnelle X n'ont pas été mesurées)?

quelle attitude adopter avec les observations « suspectes », c'est-à-dire les observations qui contrastent avec les autres?

La réponse à ces questions est donnée par divers tests statistiques: tests d'indépendance et de stationnarité des observations, d'homogénéité de deux ou plusieurs échantillons, de symétrie de la répartition, d'ajustement, etc. Ce chapitre est consacré à la description de ces tests.

11.1. Test d'ajustement du modèle de loi choisi avec les données initiales

Supposons qu'on ait postulé que la suite d'observations (11.1) est un échantillon extrait d'une population générale de fonction de répartition modèle $F_{\text{mod}}(X; \theta^{(1)}, \ldots, \theta^{(s)})$, où la forme de la fonction F_{mod} (c'est-à-dire le type du modèle) est connue, et les paramètres $\theta^{(1)}, \ldots, \theta^{(s)}$ peuvent être connus ou inconnus.

Les tests décrits dans ce paragraphe sont destinés à vérifier l'hypothèse

$$H_0: F_\xi(x) = F_{\text{mod}}(x;\ \theta^{(1)}, \ldots, \theta^{(s)}) \tag{11.2}$$

en utilisant les diverses notions de distances entre la fonction de répartition empirique analysée $\hat{F}^{(n)}(x)$ (définie sur le vu de l'échantillon (11.1), cf. § 5.5) et la fonction de répartition hypothétique $F_{\text{mod}}(x;\ \theta^{(1)}, \ldots, \theta^{(s)})$.

11.1.1. Test du χ^2 de Pearson. Le test du χ^2 permet de vérifier l'hypothèse (11.2) dans les cas où les valeurs des paramètres $\theta^{(1)}, \ldots, \theta^{(s)}$ de la fonction de répartition hypothétique sont inconnues. Pour mesurer le désaccord entre la fonction de répartition empirique et la fonction modèle, ce test utilise la statistique du χ^2 introduite dans le n° 6.2.1 (cf. formule (6.20)). La procédure de test de l'hypothèse (11.2) comporte les étapes suivantes.

1. L'intervalle des valeurs de la variable aléatoire ξ est divisé en k intervalles de groupement $\Delta_1, \ldots, \Delta_k$ pas forcément égaux. Cette partition doit nécessairement remplir les conditions suivantes:

a) le nombre total k d'intervalles doit être $\geqslant 8$*);

b) chaque intervalle de groupement doit contenir au moins 7 à 10 valeurs observées de ξ, et il est souhaitable qu'il y ait à peu près le même nombre de valeurs dans chaque intervalle;

c) si l'intervalle de variation de ξ est la droite numérique tout entière (resp. une demi-droite), les deux (resp. l'un des deux) intervalles de groupement extrêmes seront (resp. sera) des (resp. une) demi-droites.

2. Les estimations statistiques $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(s)}$ des paramètres inconnus $\theta^{(1)}, \ldots, \theta^{(s)}$ sont construites (cf. chap. 8) sur le vu des données empiriques $x_1, x_2, \ldots, x_n$. On admet généralement que la procédure la plus correcte de calcul des estimations $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(s)}$ est celle qui fait intervenir les données groupées.

3. On compte les nombres ν_i de valeurs qui sont tombées dans chaque intervalle de groupement Δ_i et on calcule les probabilités des événements $\xi \in \Delta_i$, c'est-à-dire les probabilités

$$p_i = F_{\text{mod}}(\tilde{x}_i^0,\ \hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(s)}) - F_{\text{mod}}(\tilde{x}_{i-1}^0;\ \hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(s)})$$

de tomber dans les intervalles Δ_i ($\tilde{x}_{i-1}^0$ et $\tilde{x}_i^0$ sont les extrémités gauche et droite du i-ième intervalle de groupement).

---

*) On admet que le nombre s des paramètres est $\leqslant 7$ (en pratique, $s \leqslant 3$). Si l'on renonce à cette hypothèse, il faut remplacer la condition a) par une condition plus forte: $k \geqslant \max(8, s+1)$.

4. On calcule la valeur de la statistique critique $\chi^2\ (k-s-1)$ à l'aide de la formule

$$\chi^2(k-s-1)=\sum_{i=1}^{k}\frac{(\nu_i-np_i)^2}{np_i}.$$

Ensuite, dans le tableau 2.2a [27], on trouve le point de pourcentage $100\ (1-\alpha/2)$ de la répartition du $\chi^2_{1-\alpha/2}\ (k-s-1)$ et le point de pourcentage $100\ \alpha/2$ de la répartition du $\chi^2_{\alpha/2}\ (k-s-1)$ (comme toujours α est le niveau qui est donné *a priori*).

Si

$$\chi^2_{1-\alpha/2}(k-s-1)\leqslant\chi^2(k-s-1)<\chi^2_{\alpha/2}(k-s-1),$$

l'hypothèse que la variable aléatoire ξ suit bien la loi de fonction de répartition F_{mod} est acceptée.

L'inégalité

$$\chi^2(k-s-1)\geqslant\chi^2_{\alpha/2}(k-s-1)$$

exprime un profond désaccord entre la fonction de répartition empirique et la fonction de répartition hypothétique $F_{\text{mod}}(x)$.

Le cas

$$\chi^2(k-s-1)\leqslant\chi^2_{1-\alpha/2}(k-s-1)$$

implique un complément d'étude *).

Si, par exemple, on teste l'hypothèse de normalité, la fonction de répartition hypothétique sera de la forme:

$$F_{\text{mod}}(x;\ a,\ \sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(u-a)^2}{2\sigma^2}}\,du,$$

---

*) Le rejet de l'hypothèse dans le cas où les valeurs de la statistique $\gamma^{(n)}$ sont trop petites semble à première vue contraire au bon sens. En effet, la statistique $\gamma^{(n)}$ caractérise le degré du désaccord entre la fonction de répartition empirique de la variable aléatoire ξ et la fonction de répartition hypothétique F_{mod}: plus $\gamma^{(n)}$ est petite, plus ce désaccord l'est. L'idéal semble-t-il serait que $\gamma^{(n)}=0$. Qu'on ne s'y trompe pas: $\gamma^{(n)}$ mesure certes le désaccord entre la fonction de répartition empirique et la fonction hypothétique, mais c'est une mesure aléatoire, c'est-à-dire sujette obligatoirement à une dispersion incontrôlable. De ce point de vue sont aussi peu plausibles les trop petites comme les trop grandes valeurs de $\gamma^{(n)}$. Qu'expriment les trop petites valeurs de $\gamma^{(n)}$? Plusieurs choses. Que la fonction de répartition hypothétique est sans doute mal choisie: elle dépend d'un nombre artificiellement élevé de paramètres. Que la technique d'échantillonnage est incorrecte ou subjective. Que l'analyste a essayé d'« ajuster » les données empiriques au résultat espéré, etc.

et les estimations $\hat{a}$ et $\hat{\sigma}^2$ des paramètres inconnus a et σ^2 seront

$$\hat{a} = \frac{1}{n} \sum_{i=1}^{k} \nu_i x_i^0;$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{k} \nu_i (x_i^0 - \hat{a})^2$$

(par x_i^0 on désigne comme toujours le milieu de l'intervalle Δ_i).

Les valeurs de $F(\tilde{x}_j^0; \hat{a}, \hat{\sigma}^2)$ nécessaires au calcul des probabilités p_i sont accessibles, par exemple, dans la table 1.1 [27] des valeurs de la fonction de répartition de la loi normale réduite compte tenu de la relation $F(x; a, \sigma^2) = F\left(\frac{x-a}{\sigma}; 0, 1\right)$. Le nombre de degrés de liberté de la loi du χ^2 (nous aurons besoin des points de pourcentage de cette loi) est égal ici à $k - 3$, où k est le nombre d'intervalles de groupement.

11.1.2. Test de la normalité d'une répartition d'après son coefficient de dissymétrie, son coefficient d'aplatissement et ses écarts moyens. Pour vérifier approximativement qu'une variable aléatoire ξ est normale sur le vu des observations $x_1, \ldots, x_n$, on peut utiliser certaines propriétés caractéristiques de la loi normale. On sait, en particulier, que pour la loi normale les coefficients de dissymétrie β_1 et d'aplatissement β_2 sont nuls, que l'écart absolu moyen $\theta = \mathbf{E} \mid \xi - a \mid$ est relié à l'écart-type $\sigma = \sqrt{\mathbf{E}(\xi - a)^2}$ par la relation $\theta/\sigma = \sqrt{2/\pi}$ (cf. n° 6.1.5).

Vu que nous n'avons affaire qu'aux valeurs *empiriques approchées* $\hat{\beta}_1(n)$ du coefficient de dissymétrie, $\hat{\beta}_2(n)$ du coefficient d'aplatissement, $\hat{\theta}_n$ de l'écart absolu moyen et $s^2(n)$ de l'écart-type, valeurs qui sont sujettes inévitablement à une dispersion incontrôlable, nous ne pouvons exiger la réalisation rigoureuse des relations

$$\hat{\beta}_1(u) = \hat{\beta}_2(n) = 0; \tag{11.3}$$

$$d_n = \hat{\theta}_n / s(n) = \sqrt{\frac{2}{\pi}}. \tag{11.4}$$

Mais si l'on postule la normalité de la loi de la variable aléatoire étudiée ξ, on peut établir et tabuler les lois de probabilité de $\beta_1(n)$, $\beta_2(n)$, d_n (ou de certaines de leurs combinaisons qui serviraient nos objectifs) et, partant, déterminer les écarts « tolérables » par rapport à (11.3) et (11.4) ou à des relations déduites à partir d'elles. Considérons trois procédés de vérification de la normalité basée sur les propriétés de β_1, β_2, θ et σ indiquées ci-dessus.

1. Pour vérifier la normalité d'après le coefficient de dissymétrie $\hat{\beta}_1(n)$ et le rapport $d_n = \hat{\theta}_n/s(n)$ des écarts moyens dans les cas où la *taille des échantillons n'est pas très élevée* ($n \geqslant 25$), on aura intérêt à utiliser les tableaux pour le calcul (en fonction des paramètres Q et n donnés) des valeurs exactes des points de pourcentage $100\,Q$ de la répartition des statistiques $\hat{\beta}_1(n)$ et d_n. En particulier, la procédure de vérification de la normalité est la suivante: on calcule $\hat{\beta}_1(n)$ avec la formule (5.33') et d_n avec la formule

$$d_n = \frac{\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}(n)|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}(n))^2}}.$$

En prenant Q_1 égal à 0,01 ou 0,05 et connaissant la taille n de l'échantillon, on détermine sur le tableau 4.7b de [27] la valeur du point $\gamma_{1\cdot Q_1}(n)$ de pourcentage $100\,Q_1$; en prenant ensuite Q_2 égal à 0,01, 0,05 ou 0,10, on trouve sur le tableau 4.7a de [27] les valeurs du point $d_{n\cdot Q_2}$ de pourcentage $100\,Q_2$ et du point $d_{n\cdot 1-Q_2}$ de pourcentage $100\,(1-Q_2)$; si l'une au moins des inégalités

$$\begin{cases} |\hat{\beta}_1(n)| < \gamma_{1\cdot Q_1}(n); \\ d_{n\cdot 1-Q_2} < d_n < d_{n\cdot Q_2} \end{cases}$$

est violée, l'hypothèse de normalité est rejetée au seuil de signification α satisfaisant la double inégalité *):

$$2\max(Q_1, Q_2) < \alpha < 2(Q_1 + Q_2) - 2Q_1Q_2. \qquad (11.5)$$

2. Si l'on vérifie la normalité d'une loi d'après les coefficients de dissymétrie $\hat{\beta}_1(n)$ et d'aplatissement $\hat{\beta}_2(n)$ pour des *échantillons de*

---

*) Quand on aura à tester une hypothèse H_0 à l'aide d'une statistique critique *à deux dimensions* $\gamma^{(n)} = (\gamma_1^{(n)}, \gamma_2^{(n)})$ dont sont connues (tabulées) les répartitions *marginales* de ses seules composantes, on se servira des relations et estimations suivantes. Soit A un événement lié à la variable $\gamma_1^{(n)}$ et exprimant que le test de l'hypothèse H_0 a fourni un résultat positif par rapport à la composante $\gamma_1^{(n)}$. Soit B le même événement pour la deuxième composante $\gamma_2^{(n)}$. Les tableaux des répartitions marginales de $\gamma_1^{(n)}$ et $\gamma_2^{(n)}$ permettent de définir A et B à l'aide de valeurs données peu élevées de Q_1 et Q_2, de telle sorte que $P(A) = 1 - Q_1$ et $P(B) = 1 - Q_2$. On admettra par ailleurs que $P\{A \mid B\} > P\{A\}$ et $P\{B \mid A\} > P\{B\}$, ce qui, en principe, ne contredit pas la relation logique et stochastique liant $\gamma_1^{(n)}$ et $\gamma_2^{(n)}$. Dans ces conditions, si l'on rejette l'hypothèse H_0 dans le cas où *l'un au moins* des événements A et B serait faux, on obtient sans peine l'encadrement suivant pour le seuil de signification α: $\max(Q_1, Q_2) < \alpha < (Q_1 + Q_2) - Q_1Q_2$.

taille modérément élevée ($n \geqslant 50$), on peut en plus de la table 4.7b de [27] utiliser les tables des points de pourcentage de la répartition du coefficient d'aplatissement empirique. A cet effet, il faut calculer le coefficient de dissymétrie empirique $\hat{\beta}_1(n)$ avec la formule (5.33'); prendre $Q_1 = 0{,}01$ ou $0{,}05$ et trouver la valeur du point $\gamma_{1\cdot Q_1}(n)$ de pourcentage $100Q_1$ dans la table; calculer le coefficient d'aplatissement empirique $\hat{\beta}_2(n)$ avec la formule (5.34'); prendre $Q_2 = 0{,}01$ ou $0{,}05$ pour trouver les valeurs du point $\gamma_{2\cdot Q_2}(n)$ de pourcentage $100Q_2$ et du point $\gamma_{2\cdot 1-Q_2}(n)$ de pourcentage $100(1 - Q_2)$ dans la table 4.7c de [27]. Si l'une au moins des inégalités

$$\begin{cases} |\hat{\beta}_1(n)| < \gamma_{1\cdot Q_1}(n); \\ \gamma_{2\cdot 1-Q_2}(n) < \hat{\beta}_2(n) < \gamma_{2\cdot Q_2}(n) \end{cases}$$

est mise en défaut, l'hypothèse de normalité est à rejeter à un seuil de signification α satisfaisant la double inégalité (11.5).

3. Pour vérifier la normalité d'une loi d'après les coefficients de dissymétrie et d'aplatissement pour des *échantillons de taille assez élevée* (n est de l'ordre de 10^3), il est préférable de se servir de la normalité approximative (asymptotiquement par rapport à n) de la répartition des coefficients de dissymétrie $\hat{\beta}_1(n)$ et d'aplatissement $\hat{\beta}_2(n)$ empiriques. Signalons toutefois que la loi de $\hat{\beta}_2(n)$ converge très lentement vers la loi normale: même si n est de l'ordre de plusieurs centaines, on peut mettre en évidence l'important écart entre les valeurs *exactes* des points de pourcentage de la loi de $\hat{\beta}_2(n)$, trouvées dans la table 4.7c de [27], et les valeurs *approchées* des mêmes points de pourcentage, calculées à l'aide de l'approximation normale de paramètres $\left(-\frac{6}{n+1}, \sqrt{\frac{24}{n}\left(1 - \frac{225}{15n + 124}\right)}\right)$ de la répartition de $\hat{\beta}_2(n)$. Ceci est, en particulier, la cause des trop grossiers résultats que l'on obtient lorsque dans la vérification de la normalité d'après $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$ on se sert, pour des n relativement peu élevés, de la normalité approchée pour trouver les points de pourcentage des répartitions de $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$.

Si la taille de l'échantillon est assez grande (de l'ordre de 10^3), on peut mettre en œuvre la procédure suivante de vérification de la normalité de la variable aléatoire ξ sur le vu des observations $x_1, \ldots, x_n$:

on calcule les coefficients empiriques $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$ avec les formules (5.33') et (5.34');

on détermine les écarts-types des coefficients $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$ avec les formules (cf. [76]):

$$\sigma_{\hat{\beta}_1} = \sqrt{V\hat{\beta}_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}\left(1 - \frac{12}{2n+7}\right)};$$

$$\sigma_{\hat{\beta}_2} = \sqrt{V\hat{\beta}_2} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}} \approx \sqrt{\frac{24}{n}\left(1 - \frac{225}{15n+124}\right)};$$

pour q donné proche de l'unité (par exemple $q = 0{,}95$) on trouve sur la table la valeur du quantile u_q de la loi normale;

si l'une au moins des inégalités

$$\begin{cases} |\hat{\beta}_1(n)| < u_q \cdot \sigma_{\hat{\beta}_1}; \\ \left|\hat{\beta}_2(n) + \frac{6}{n+1}\right| < u_q \cdot \sigma_{\hat{\beta}_2} \end{cases}$$

est violée *), l'hypothèse de la normalité de la variable aléatoire ξ est rejetée à un seuil de signification α satisfaisant (11.5), où $Q_1 = Q_2 = 1 - q$, c'est-à-dire que $2(1-q) < \alpha < 4(1-q) - 2(1-q)^2$.

Ce test de normalité admet plusieurs modifications différentes. On peut, en particulier, utiliser une statistique à une dimension de la forme

$$\varkappa^2 = c_{1n}\hat{\beta}_1^2 + c_{2n}\hat{\beta}_2^2,$$

où c_{1n} et c_{2n} sont des coefficients > 0 dépendant de n. C'est ainsi qu'on est conduit à la statistique $\varkappa^2$, si, pour mesure du désaccord entre la densité de probabilité empirique $\hat{f}(x)$ et la densité théorique de la loi normale $f(x;\ a,\ \sigma^2)$, on prend l'analogue continu de la statistique χ^2 dans le test

$$\tilde{\chi}^2 = n \int_{-\infty}^{\infty} \frac{|\hat{f}_n(t) - f(t)|^2}{f(t)}\, dt,$$

où $t = (x - a)/\sigma$. Pour obtenir une expression approchée de χ^2 à partir de $\tilde{\chi}^2$, il faut développer $\hat{f}_n(t)$ en série de Edgeworth (cf. [40, n° 17.7]) et se limiter aux trois premiers termes. Mais les lois exactes (pour chaque n fixe) des statistiques χ^2 n'ont pas été calculées, quant aux approximations basées sur la normalité asymptotique de $\hat{\beta}_1(n)$ et $\hat{\beta}_2(n)$, elles sont trop grossières.

*) La deuxième de ces inégalités prend en considération le *biais* de l'estimation $\hat{\beta}_2(n)$: si $x_1, \ldots, x_n$ est un échantillon normal, alors $E\hat{\beta}_2(n) = -\frac{6}{n+1}$ (et non pas $E\hat{\beta}_2(n) = \beta_2 = 0$).

11.1.3. Test de Kolmogorov-Smirnov et son application à la construction de la région de confiance pour la fonction de répartition inconnue. Le test de Kolmogorov-Smirnov permet d'éprouver l'hypothèse (11.2) lorsque la fonction modèle $F_{\text{mod}}(x) = F_0(x)$ *est entièrement connue*, c'est-à-dire ne dépend pas de paramètres inconnus.

Statistiques du test de Kolmogorov-Smirnov et leurs répartitions. Soit $F^{(n)}(x)$ une fonction de répartition empirique. Introduisons les mesures suivantes du désaccord entre $F^{(n)}(x)$ et $F_0(x)$:

$$\begin{aligned} D_n &= \sup_{x \in R^1} | F^{(n)}(x) - F_0(x) | ; \\ D_n^+ &= \sup_{x \in R^1} (F^{(n)}(x) - F_0(x)) ; \\ D_n^- &= \sup_{x \in R^1} (F_0(x) - F^{(n)}(x)). \end{aligned} \tag{11.6}$$

Les statistiques $\sqrt{n} D_n$ et $\sqrt{n} D_n^-$ sont les statistiques respectives des tests de Kolmogorov et de Smirnov. Ceci étant,

$$D_n = \max (D_n^+, D_n^-).$$

Pour utiliser pratiquement le test de Kolmogorov-Smirnov, on met les statistiques D_n, D_n^+ et D_n^- sous la forme:

$$\begin{aligned} D_n^+ &= \max_{1 \leqslant i \leqslant n} \left(\frac{i}{n} - t_i \right) ; \\ D_n^- &= \max_{1 \leqslant i \leqslant n} \left(t_i - \frac{i-1}{n} \right) ; \\ D_n &= \max_{1 \leqslant i \leqslant n} (D_n^+, D_n^-), \end{aligned} \tag{11.7}$$

où $t_i = F_0(x_{(i)})$, c'est-à-dire que t_i est la valeur prise par la fonction de répartition hypothétique au point i de la série variationnelle.

Les fonctions de répartition exactes des statistiques D_n, D_n^+ et D_n^- sont connues [27], [122]. Exhibons seulement celle de D_n^+:

$$P\{D_n^+ \leqslant d\} =$$

$$= \begin{cases} 0, & d \leqslant 0 ; \\ 1 - \sum\limits_{j=0}^{[n-nd]} C_j^n \left(\frac{j}{n} + d \right)^{j-1} \left(1 - \frac{j}{n} - d \right)^{n-j}, & 0 < d < 1 ; \\ 1, & d \geqslant 1, \end{cases}$$

où $[n-nd]$ est la partie entière de $n - nd$.

Dans la pratique, on se contente des fonctions de répartition limites des statistiques $\sqrt{n}D_n$ et $\sqrt{n}\,D_n^+$:

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n < \lambda) = K(\lambda);$$

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n^+ < \lambda) = S(\lambda),$$

où

$$K(\lambda) = \begin{cases} 1 - 2\sum\limits_{k=1}^{\infty} (-1)^k e^{-2k^2\lambda^2}, & \lambda > 0; \\ 0, & \lambda < 0; \end{cases} \tag{11.8}$$

$$S(\lambda) = \begin{cases} 1 - e^{-2\lambda^2}, & \lambda \geqslant 0; \\ 0, & \lambda < 0. \end{cases} \tag{11.9}$$

La fonction de répartition limite de la statistique $\sqrt{n}\,D_n^-$ est confondue avec $S(\lambda)$.

S'agissant de la fonction de répartition limite de la statistique de Kolmogorov $\sqrt{n}\,D_n$, on sait [77] que si $K_n(\lambda)$ est la fonction de répartition exacte de la statistique $\sqrt{n}\,D_n$, alors $K_n(\lambda) > K(\lambda)$ pour tout n, et l'erreur maximale commise, en remplaçant la fonction de répartition exacte par la fonction de répartition limite pour $n \geqslant 60$, est de l'ordre de $0{,}8 \cdot 10^{-2}$.

Région de confiance de la fonction de répartition. Vu que la fonction de répartition de la statistique de Kolmogorov D_n est indépendante de la répartition inconnue de la population générale et que la distance entre $F^{(n)}(x)$ et $F_{\text{mod}}(x)$ est mesurée à l'aide de l'écart maximal, on peut inverser le test et utiliser D_n pour déterminer la région de confiance de la fonction de répartition continue [77]. Pour toute fonction de répartition $F_{\text{mod}}(x)$, on a

$$P\{F^{(n)}(x) - d_\alpha \leqslant F_{\text{mod}}(x) \leqslant F^{(n)}(x) + d_\alpha,\ \forall x\} = 1 - \alpha, \tag{11.10}$$

où d_α est la valeur critique de D_n correspondant au seuil de signification α.

La région de confiance est donc la bande $\pm\, d_\alpha$ le long de la fonction de répartition empirique $F^{(n)}(x)$ et la fonction théorique $F_{\text{mod}}(x)$ est entièrement contenue à l'intérieur de cette bande au seuil de $1 - \alpha$.

En se servant de ce résultat, on peut établir des estimations de la taille de l'échantillon, nécessaires pour approcher la fonction de répartition avec la précision demandée.

Pour $\alpha \leqslant 0{,}2$ et $n \geqslant 80$, on a

$$d_\alpha \approx \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}}\Big/\sqrt{n}. \tag{11.11}$$

Posons, par exemple, $\alpha = 0{,}05$. Pour un échantillon de taille n, l'écart entre la fonction de répartition théorique et la fonction de répartition empirique est $\leqslant \Delta = 0{,}61/\sqrt{n}$ au seuil de 0,95. Pour $n = 100$, on a $\Delta = 0{,}061$.

11.1.4. Test du ω^2 (Cramer-Mises-Smirnov). La statistique de ce test est la distance en moyenne quadratique entre les fonctions de répartition théorique et empirique:

$$W_n^2 = n\omega_n^2 = n \int (F_0(x) - F^{(n)}(x))^2 \, dF_0(x). \qquad (11.12)$$

Cette statistique est utilisée dans les applications sous la forme suivante:

$$W_n^2 = \sum_{i=1}^{n} \left(F_0(x_{(i)}) - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}. \qquad (11.13)$$

Si l'hypothèse H_0 est vraie, la loi de la statistique $n\omega_n^2$ converge pour $n \to \infty$ vers une loi limite de fonction de répartition $a_1(x)$. La table de cette répartition est donnée dans [27].

11.1.5. Modifications des statistiques des tests de Kolmogorov-Smirnov et du ω^2 pour les échantillons de petite taille. Les fonctions de répartition des statistiques D_n^+, D_n^-, D_n et W_n^2 tendent rapidement vers les fonctions de répartition limites lorsque la taille n de l'échantillon croît. Si la taille n de l'échantillon est petite, le désaccord entre la fonction de répartition limite et la fonction de répartition correspondant à n fini peut être considérable, ce qui nécessite l'usage

Tableau 11.1

Statistique primitive	Statistique modifiée
	Aile droite de la répartition
$\sqrt{n}\, D_n^+$	$\tilde{D}_n^+ = D_n^+ (\sqrt{n} + 0{,}12 + 0{,}11/\sqrt{n})$
$\sqrt{n}\, D_n^-$	$\tilde{D}_n^- = D_n^- (\sqrt{n} + 0{,}12 + 0{,}11/\sqrt{n})$
$\sqrt{n}\, D_n$	$\tilde{D}_n = D_n (\sqrt{n} + 0{,}12 + 0{,}11/\sqrt{n})$
W_n^2	$W^2 = (W_n^2 - 0{,}4/n + 0{,}6/n^2)(1{,}0 + 1{,}0/n)$
	Aile gauche de la répartition
$\sqrt{n}\, D_n$	$\tilde{D}_n = D_n (\sqrt{n} + 0{,}275 - 0{,}04\sqrt{n})$
W_n^2	$\tilde{W}_n^2 = (W_n^2 - 0{,}03/n)(1{,}0 + 0{,}5/n)$

des fonctions de répartition exactes et des tables respectives. Mais par une transformation peu compliquée des statistiques initiales, on peut obtenir des statistiques $\widetilde{D}_n^+$, $\widetilde{D}_n^-$, $\widetilde{D}_n$ et $\widetilde{W}_n^2$ dont les fonctions de répartitions concordent bien mieux avec les fonctions de répartition limites pour n petit [128]. Ces modifications sont données dans le tableau 11.1.

Si l'on se sert de tests basés sur les modifications $\widetilde{D}_n^+$, $\widetilde{D}_n^-$, $\widetilde{D}_n$ et $\widetilde{W}_n^2$, on ne pourra faire intervenir que les fonctions de répartition limites de $\sqrt{n}\,D_n^+$, $\sqrt{n}\,D_n^-$, $\sqrt{n}\,D_n$ et W_n^2 respectivement.

11.1.6. Technique statistique de réalisation pratique des tests non paramétriques. La méthode d'application des tests non paramétriques envisagés est dans les grandes lignes la même. Soient K_n et $P(x)$ la valeur de la statistique d'un test et sa fonction de répartition (exacte ou limite). Pour éprouver une hypothèse, il faut

1) ranger les éléments de l'échantillon par ordre de grandeur croissante;

2) calculer la valeur de la statistique associée au test K_n (ou de la modification correspondante);

3) calculer la probabilité $\alpha = 1 - P(K_n)$ ou comparer la valeur K_n aux points de pourcentage de la répartition correspondante.

Si la probabilité $\alpha = 1 - P(K_n)$ est petite (K_n est grand), c'est qu'un événement peu vraisemblable s'est produit et l'hypothèse que la série d'observations $(x_1, \ldots, x_n)$ suit la loi postulée $F_0(x)$ est rejetée, quant au désaccord entre $F^{(n)}(x)$ et $F_0(x)$, il ne peut être expliqué par le caractère aléatoire de l'expérience. L'hypothèse H_0 est parfois rejetée pour des valeurs K_n trop petites et, par suite, à des seuils $\alpha = 1 - P(K_n)$ proches de 1*).

Exemple 11.1. Appliquons les testsdu D_n et du ω_n^2 pour vérifier la qualité d'un générateur de nombres uniformément répartis sur l'intervalle (0, 36).

Supposons que le générateur ait délivré les nombres: 23 18 1 16 2 3 20 4 7 0. Les résultats des calculs sont représentés sur le tableau 11.2.

Les formules (11.7) et (11.13) nous donnent: $D_{10} = 0{,}4056$; $\sqrt{n}D_{10} = 1{,}2826$; $\widetilde{D}_{10} = 1{,}3454$; $W_{10}^2 = 0{,}6209$; $\widetilde{W}_{10}^2 = 0{,}5335$. Respectivement, en se servant des tables des fonctions de répartitions limites pour $\sqrt{n}\,D_n$ et W^2 de [27], on trouve pour les modifications $\widetilde{D}_{10}$ et $\widetilde{W}_{10}$:

$$P(\widetilde{D}_{10} > 1{,}3454) \approx 0{,}05; \quad P(\widetilde{W}_{10}^2 > 0{,}5335) \approx 0{,}048.$$

Nous n'avons donc aucune raison de rejeter l'hypothèse concernant l'uniformité de la répartition des nombres délivrés par le générateur.

On peut aboutir au même résultat en comparant les valeurs de

*) Voir note de la page 299.

Tableau 11.2

i	$x_{(i)}$	$F^{(n)}(x) = \frac{i}{n}$	$F_0(x) = t_i$	$\frac{i}{n} - t_i$	$t_i - \frac{2i-1}{2n}$
1	0	0,1	0	0,1	—0,05
2	1	0,2	0,0278	0,1722	—0,1222
3	2	0,3	0,0556	0,2444	—0,1944
4	3	0,4	0,0833	0,2167	—0,1667
5	4	0,5	0,1111	0,3889	—0,3389
6	7	0,6	0,1944	0,4056	—0,3556
7	16	0,7	0,4444	0,2556	—0,2056
8	18	0,8	0,5	0,3000	—0,2500
9	20	0,9	0,5556	0,3444	—0,2944
10	23	1,0	0,6389	0,3611	—0,3111

D_{10} et W_{10}^2 aux points de pourcentage des répartitions exactes respectives. Ainsi, pour $\alpha = 0{,}05$, la valeur critique $d_{0,05}^{(10)} = 0{,}4092$ et, par suite, $D_{10} < d_{0,05}^{(10)}$.

11.1.7. Application des tests de Kolmogorov et du ω^2 pour vérifier la normalité d'une fonction de répartition dépendant de paramètres inconnus. Si la fonction de répartition hypothétique est connue aux paramètres près et que ces derniers soient estimés sur le vu de l'échantillon, les répartitions limites des statistiques D_n, D_n^+, D_n^-, ω_n^2 ne seront déjà plus « distribution free ». C'est là que réside leur handicap par rapport au test du χ^2 (cf. n° 11.1.1). Mais si l'estimation concerne les paramètres de localisation et d'échelle, la répartition ne dépendra que de la forme de la fonction de répartition $F(x)$ et pas de ses paramètres [77]. L'indépendance des répartitions limites par rapport aux paramètres de localisation ét d'échelle permet de construire, par exemple sur la base des statistiques indiquées, des tests vérifiant la normalité de la répartition.

Vérification de la normalité d'une répartition par le test du ω^2. Supposons qu'on nous demande de vérifier si une fonction de répartition inconnue est normale. Trois cas peuvent se présenter.

1. On connaît la variance σ et pas la moyenne m. La moyenne m est estimée par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2. On connaît la moyenne m et pas la variance σ. La quantité σ^2 est estimée par la statistique

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2.$$

3. On ne connaît ni la moyenne m ni la variance σ. Les valeurs de m et de σ^2 sont estimées respectivement par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Des tables des répartitions limites pour la statistique W^2 ont été dressées pour chacune de ces situations dans le travail [92].

Tableau 11.3 [44]

$1-\alpha$	Paramètres inconnus			$1-\alpha$	Paramètres inconnus		
	m	σ^2	m et σ^2		m	σ^2	m et σ^2
0,50	0,0627	0,1017	0,0509	0,95	0,1653	0,4418	0,1260
0,60	0,0729	0,1272	0,0585	0,99	0,2380	0,7245	0,1788
0,70	0,0857	0,1635	0,0680	0,995	0,2698	0,8506	0,2018
0,80	0,1036	0,2200	0,0811	0,999	0,3443	1,1490	0,2559
0,85	0,1164	0,2631	0,0904	0,9999	0,4527	1,5860	0,3344
0,90	0,1344	0,3270	0,1035				

Tableau 11.4 [44]

n	0,10	0,05	0,01
4	0,352	0,381	0,417
5	0,317	0,337	0,405
6	0,294	0,319	0,364
7	0,276	0,300	0,348
8	0,261	0,285	0,331
9	0,249	0,271	0,311
10	0,239	0,258	0,294
11	0,230	0,249	0,284
12	0,223	0,242	0,275
13	0,214	0,234	0,268
14	0,207	0,227	0,261
15	0,201	0,220	0,257
16	0,195	0,213	0,250
17	0,189	0,206	0,245
18	0,184	0,200	0,239
19	0,179	0,195	0,235
20	0,174	0,190	0,231
25	0,165	0,180	0,203
30	0,144	0,161	0,187
$n > 30$	$\frac{0,805}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$

Le tableau 11.3 représente les valeurs critiques d'une statistique de type ω^2 à un seuil de signification α compris entre 0,5 et 0,0001. L'approximation de l'aile droite des répartitions correspondantes est étudiée dans le n° 12.2.7.

Vérification de la normalité d'une répartition à l'aide de la statistique de Kolmogorov. La fonction de répartition de la statistique de Kolmogorov dans le cas où la variance et la moyenne sont inconnues a été étudiée et tabulée dans [87]. Les valeurs critiques de la statistique D_n pour $\alpha = 0{,}1$; 0,05 ; 0,01 sont représentées sur le tableau 11.4 (les paramètres m et σ^2 sont tous deux inconnus).

11.2. Test des hypothèses d'homogénéité et de symétrie d'une répartition

Soient donnés $k \geqslant 2$ échantillons indépendants constitués respectivement de $n_1, \ldots, n_k$ observations indépendantes :

$$\begin{array}{c} x_1, \ldots, x_{n_1}; \\ y_1, \ldots, y_{n_2}; \\ \ldots\ldots\ldots\ldots \\ z_1, \ldots, z_{n_k}. \end{array}$$

L'*hypothèse d'homogénéité* exprime que les populations générales d'où sont extraits les échantillons sont identiques et, par conséquent, ont les mêmes fonctions de répartition :

$$H_0 : F_1(x) = F_2(x) = \ldots = F_k(x) = F(x), \qquad (11.14)$$

où $F_i(x)$ est la fonction de répartition de la i-ième population générale. Le cas particulier le plus fréquent dans les applications est celui où $k = 2$.

Le test de l'*hypothèse de symétrie* consiste dans ce paragraphe à s'assurer que la fonction de répartition est symétrique par rapport à un point donné μ :

$$H_0 : F(x + \mu) = 1 - F(\mu - x). \qquad (11.15)$$

Si la répartition admet une densité, l'hypothèse H_0 revient à vérifier que cette densité est symétrique (par rapport au point μ).

11.2.1. Tests d'homogénéité basés sur les fonctions de répartition empiriques. Traitons le cas de deux échantillons à une dimension ($k = 2$). Soient $x_{(1)}, \ldots, x_{(n_1)}$ et $y_{(1)}, \ldots, y_{(n_2)}$ les séries variationnelles associées au premier et au deuxième échantillon. Nous pouvons définir maintenant deux fonctions de répartition empiriques $F_1^{(n_1)}(x)$ et $F_2^{(n_2)}(x)$. Comme dans le n° 11.1.4, introduisons les

statistiques suivantes:

$$D^+_{n_1, n_2} = \sup_x (F_1^{(n_1)}(x) - F_2^{(n_2)}(x)) ;$$

$$D^-_{n_1, n_2} = \sup_x (F_2^{(n_2)}(x) - F_1^{(n_1)}(x)) ; \tag{11.16}$$

$$D_{n_1, n_2} = \max (D^-_{n_1, n_2}, D^+_{n_1, n_2}) = \sup_x | F_1^{(n_1)}(x) - F_2^{(n_2)}(x) | .$$

Si l'hypothèse de base est vraie, les statistiques $D^+_{n_1,n_2}$ et $D^-_{n_1,n_2}$ ont la même fonction de répartition, donc on n'étudiera que l'une d'elles, disons, $D^+_{n_1,n_2}$. Sans nuire à la généralité, on peut admettre que $n_2 \leqslant n_1$. Supposons maintenant que les fonctions limites $F_1(x)$ et $F_2(x)$ sont continues et que l'hypothèse H_0 est vraie. Supposons, enfin, que $n_2 \to \infty$ et que $n_0 = \frac{n_1 n_2}{n_1 + n_2}$. Sous ces conditions, les variables aléatoires $\sqrt{n_0}\, D^+_{n_1,n_2}$, et $\sqrt{n_0} D_{n_1,n_2}$ ont les mêmes fonctions de répartition limites $S(\lambda)$ et $K(\lambda)$ que leurs analogues (11.6). Les fonctions de répartition des stastistiques (11.16) pour $n_1 = n_2 = n$ finis ont été établies par B. Gnédenko et V. Koroliouk [61] (pour des valeurs rationnelles de $\lambda = k/n$):

$$P\left(D^+_{(n)} \leqslant \frac{k}{n}\right) = 1 - \frac{C_{2n}^{n+k+1}}{C_{2n}^n} ; \tag{11.17}$$

$$P\left(D_{(n)} \leqslant \frac{k}{n}\right) = \sum_{i=-r}^{r} (-1)^i \frac{C_{2n}^{n+i(k+1)}}{C_{2n}^n}, \quad r = \left[\frac{n}{k+1}\right]. \tag{11.18}$$

Les tables des répartitions exactes des statistiques $D^+_{n_1, n_2}$ et D_{n_1, n_2} dans le cas général sont accessibles dans [27].

En pratique, pour alléger les calculs on peut déterminer les quantités $D^+_{n_1, n_2}$ et D_{n_1, n_2} à l'aide des formules:

$$D^+_{n_1, n_2} = \max_{1 \leqslant j \leqslant n_2} \left(\frac{j}{n_2} - F_1^{(n_1)}(y_{(j)})\right) =$$

$$= \max_{1 \leqslant i \leqslant n_1} \left\{F_2^{(n_2)}(x_{(i)}) - \frac{i-1}{n_1}\right\} ; \tag{11.19}$$

$$D^-_{n_1 n_2} = \max_{1 \leqslant j \leqslant n_2} \left(F_1^{(n_1)}(y_{(j)}) - \frac{j-1)}{n_2}\right) =$$

$$= \max_{1 \leqslant i \leqslant n_1} \left\{\frac{i}{n_1} - F_2^{(n_2)}(x_{(i)})\right\} ; \; D_{n_1, n_2} = \max (D^+_{n_1, n_2}, D^-_{n_1, n_2}).$$

Signalons une très importante propriété des tests d'homogénéité basés sur les statistiques (11.17), (11.19): *ces tests sont convergents pour toute hypothèse concurrente de la forme*

$$H_1 : F_1(x) \neq F_2(x), \tag{11.20}$$

c'est-à-dire que, pour toute hypothèse concurrente de la forme (11.20), la probabilité de rejeter H_0 (ici (11.14) pour $k = 2$)) tend vers 1 lorsque n_1, $n_2 \to \infty$, de quelque nature que soit la différence entre $F_1(x)$ et $F_2(x)$ et quelque petite qu'elle soit.

La région critique de la statistique $\sqrt{n_0} D_{n_1, n_2}$ est le domaine des grandes valeurs *), c'est-à-dire que l'hypothèse d'homogénéité est rejetée si

$$\sqrt{n_0}\, D_{n_1, n_2} > d_\alpha(n_1, n_2),$$

où $d_\alpha(n_1, n_2)$ est un point critique de la fonction de répartition de la statistique $\sqrt{n_0} D_{n_1, n_2}$ au seuil de α.

Les grandes valeurs sont critiques également pour la statistique $D^+_{n_1, n_2}$ ($D^-_{n_1, n_2}$). Mais la statistique $D^+_{n_1, n_2}$ (tout comme $D^-_{n_1, n_2}$) peut prendre des valeurs strictement négatives. Cela exprime qu'une des fonctions de répartition empiriques est plus grande que l'autre sur l'intervalle tout entier des observations: cette conclusion est incompatible avec l'hypothèse nulle concernant l'égalité des fonctions de répartition limites pour les valeurs assez grandes de n_1 et n_2. Ainsi, pour $n_1 = n_2$, on obtient, grâce à (11.17),

$$P(D^+_{(n_1)} \leqslant 0) = 1 - C^{n_1+1}_{2n_1} / C^{n_1}_{2n_1} = 1/n_1,$$

où $D^+_{(n_1)}$ représente $D^+_{n_1, n_2}$.

Pour $n_1 \geqslant 20$, cette probabilité est $< 0{,}05$. Donc, si n_1, $n_2 > 20$, le domaine des valeurs strictement négatives des statistiques correspondantes doit être considéré comme une région critique pour l'hypothèse nulle.

Remarque. Nous avons pu obtenir la fonction de répartition des statistiques associées aux tests $D^+_{n_1, n_2}$, D_{n_1, n_2} en conjecturant la continuité des fonctions de répartition limites respectives $F_1(x)$ et $F_2(x)$. Par ailleurs, en pratique on a souvent affaire soit à des variables aléatoires discrètes, soit à des données groupées. Les formules (11.19) pour les statistiques $D^+_{n_1, n_2}$, D_{n_1, n_2} peuvent être appliquées à ce cas aussi sans aucun changement, mais le seuil de signification de l'hypothèse nulle sera inférieur au seuil donné [67]. Ceci exprime que la probabilité pour la statistique associée au test d'être supérieure au point de pourcentage 100α calculé en postulant la continuité sera $< \alpha$ et, par suite, la probabilité de rejeter l'hypothèse nulle est plus petite si les données sont groupées.

Si nous avons affaire à $k > 2$ échantillons de même taille ($n_1 = n_2 = \ldots = n_k$), on peut se servir de la généralisation suivante des statistiques:

$$D^{(k)}_{(n_1)} = \max_{i,j} \sup_x |F_i^{(n)}(x) - F_j^{(n)}(x)|. \qquad (11.21)$$

*) Au sujet des petites valeurs des statistiques critiques voir la note de la page 299.

Pour $k = 3$, on trouvera dans [44] des tables de la répartition de la statistique (11.21).

Exemple 11.2. Le tableau 11.5 représente des données (conventionnelles) concernant le salaire de $n_1 = 100$ et $n_2 = 100$ employés de deux secteurs industriels. Vérifions à l'aide des statistiques (11.16) l'hypothèse que les fonctions de répartition $F_1(x)$ et $F_2(x)$ du salaire des employés du premier et du deuxième secteur sont confondues.

Tableau 11.5

Salaire mensuel, en roubles	Nombre d'employés		Salaire mensuel, en roubles	Nombre d'employés	
	Premier secteur	Deuxième secteur		Premier secteur	Deuxième secteur
130-150	4	1	250-300	22	34
150-170	4	1	300-350	3	7
170-200	15	8	350-400	1	3
200-250	51	43	400-500	—	3

Le tableau 11.5 concerne des données groupées. Les tests basés sur les statistiques (11.16) et (11.19) sont approximatifs. En considérant les valeurs prises par les fonctions de répartition empiriques aux extrémités droites des intervalles, on obtient les données suivantes pour le calcul des statistiques critiques (tableau 11.6).

D'où

$$D^+_{(100)} = 0{,}21\,;\ D^-_{(100)} = -0{,}03\,;\ D_{(100)} = \\ = \max\,(D^+_{(100)},\ D^-_{(100)}) = 0{,}21.$$

La répartition de Gnédenko-Korolioukk (11.17) nous donne

$$P\,(D^-_{(100)} \leqslant -0{,}03) \leqslant P\,(D^-_{(100)} \leqslant 0) = 1 - \frac{99}{100} - 0{,}01,$$

autrement dit, si les fonctions de répartition $F_1(x)$ et $F_2(x)$ sont égales (i.e. l'hypothèse H_0 est vraie), la probabilité que la statisti-

Tableau 11.6

i	$x_{(i)} = y_{(i)}$	$F_1^{(n)}(x_{(i)})$	$F_2^{(n)}(x_{(i)})$	$F_1^{(n)}(x_{(i)}) - F_2^{(n)}(x_{(i)})$
1	150	0,04	0,01	0,03
2	170	0,08	0,02	0,06
3	200	0,23	0,10	0,13
4	250	0,74	0,53	0,21
5	300	0,96	0,87	0,09
6	350	0,99	0,94	0,05
7	400	1,00	0,97	0,03
8	500	1,00	1,00	0

que $D^-_{(100)}$ prendra des valeurs strictement négatives pour $n = 100$ sera inférieure à 0,01, et, par suite, l'hypothèse H_0 doit être rejetée au seuil de 0,01.

Considérons maintenant la statistique $\sqrt{n_0}D_{(n)}$:

$$n_0 = \frac{nn}{n+n} = \frac{n}{2} = 50, \quad \sqrt{n_0}\, D_{(100)} = \sqrt{50}\cdot 0{,}21 = 1{,}4849.$$

La répartition limite de la statistique de Kolmogorov $K(\lambda)$ nous donne

$$P\left(\sqrt{n_0}\, D_{(100)} \geqslant 1{,}4849\right) = 1 - K(1{,}4849) \approx 1 - 0{,}975 = 0{,}025.$$

Donc, l'hypothèse que les fonctions de répartition sont égales est aussi rejetée au seuil de 0,025. On retrouve ce résultat en comparant $\sqrt{n_0}D_{(n)}$ aux points de pourcentage de la répartition exacte. Les niveaux des tests $D^-_{(n)}$ et $\sqrt{n_0}D_{(n)}$ ne sont pas confondus dans cet exemple et l'hypothèse H_0 est moins probable par le test D^-_n que par le test $\sqrt{n_0}D_{(n)}$.

11.2.2. Test d'homogénéité du χ^2. Si les données sont groupées, chose qui est fréquente en pratique, on peut se servir d'un test du χ^2 pour vérifier l'homogénéité.

Soient donnés $k \geqslant 2$ échantillons de tailles n_i ($i = 1, \ldots, k$) et supposons que les données de chaque échantillon sont groupées dans r classes (intervalles). Désignons par ν_{ij} le nombre d'éléments du j-ième échantillon contenus dans la i-ième classe. La statistique associée à ce test est [40]:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\left(\nu_{ij} - \frac{\nu_{i.}\nu_{.j}}{n}\right)^2}{\nu_{i.}\nu_{.j}} = n \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\nu_{ij}^2}{\nu_{i.}\nu_{.j}} - 1, \quad (11.22)$$

où

$$\nu_{i.} = \sum_{l=1}^{k} \nu_{il}, \quad \nu_{.j} \sum_{l=1}^{r} \nu_{lj}, \quad n = \sum_{i=1}^{k} n_i.$$

Pour $k = 2$, la statistique (11.22) devient

$$\chi^2 = n_1 n_2 \sum_{i=1}^{r} \frac{\left(\frac{\mu_i}{n_1} - \frac{\nu_i}{n_2}\right)^2}{\mu_i + \nu_i}, \quad (11.23)$$

où μ_i, ν_i ($i = 1, \ldots, r$) sont les nombres respectifs d'éléments du premier et du deuxième échantillon contenus dans la i-ième classe. Dans le cas de l'hypothèse nulle, la quantité (11.22), (11.23)suit une loi du χ^2 à $(r-1)(k-1)$ degrés de liberté.

Exemple 11.3. Appliquons le test du χ^2 aux données de l'exemple 11.2. Les quantités nécessaires au calcul de la valeur de la statistique critique sont indiquées dans les deux dernières colonnes du tableau 11.7.

Tableau 11.7

Intervalles des salaires en roubles	Fréquences		$\mu_i + \nu_i$	$\mu_i - \nu_i$
	μ_i	ν_i		
130-150	4	1	5	3
150-170	4	1	5	3
170-200	15	8	23	7
200-250	51	43	94	8
250-300	22	34	56	—12
300-350	3	7	10	—4
350-400	1	3	4	—2
400-500	—	3	3	—3

Trouvons la valeur de la statistique associée au test :

$$\chi^2 = n_1 n_2 \sum_{i=1}^{8} \frac{\left(\frac{\mu_i}{n_1} - \frac{\nu_i}{n_2}\right)^2}{\mu_i + \nu_i} = \sum_{i=1}^{8} \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i} = 14{,}582.$$

Les tables des points de pourcentage de la loi du χ^2 à $r - 1 = 7$ degrés de liberté nous donnent $\chi^2_{0,05}(7) = 14{,}067$. Donc, l'hypothèse d'égalité peut être rejetée à un seuil $\chi^2 \approx 0{,}05$.

Exemple 11.4. Le tableau 11.8 représente la répartition des revenus d'après le recensement fait en Suède en 1930 ([40], tableau 30.6.2).

Si l'on compare la répartition des revenus des ouvriers de l'industrie dans les tranches d'âge de 40-50 ans et 50-60 ans, on trouve que $\chi^2 = 840{,}62$ avec 5 degrés de liberté, ce qui dénote une très grande différence entre les répartitions. Mais pour un groupe plus homogène d'O.S., la comparaison des répartitions des revenus pour les mêmes groupes d'âge donne $\chi^2 = 4{,}27$ $(1 - F_{\chi^2(5)}(4{,}27) \approx 0{,}51)$, de sorte que les deux derniers échantillons peuvent être considérés comme homogènes.

11.2.3. Tests de rang d'homogénéité. Les tests de rang d'homogénéité utilisent les numéros des observations dans la série variationnelle obtenue en oordonnant l'échantillon *global* de taille $n = \sum_{i=1}^{k} n_i$. Le numéro de l'observation x_i dans l'échantillon ordonné s'appelle *ordre* ou *rang* et sera désigné dans la suite par $R_{i,n}$.

Tableau 11.8

Revenus en milliers de couronnes	Tous les ouvriers et employés de l'industrie		Ouvriers spécialisés	
	Groupes d'âge		Groupes d'âge	
	40-50	50-60	40-50	50-60
0-1	7 831	7 558	71	54
1-2	26 740	20 685	430	324
2-3	35 572	24 186	1 072	894
3-4	20 009	12 280	1 609	1 202
4-6	11 527	6 776	1 178	903
6	6 919	4 222	158	112
Total:	$n_1 = 108\,598$	$n_2 = 75\,707$	$n_1 = 4\,518$	$n_2 = 3\,489$
$\chi^2 = 840{,}62$ (5 degrés de liberté)			$\chi^2 = 4{,}27$ (5 degrés de liberté)	

On étudiera les *tests de rang linéaires* dont les statistiques sont de la forme :

$$K = \sum_{i=1}^{n_j} \varphi(R_{i,n}), \tag{11.24}$$

où la sommation est étendue aux seuls éléments du premier ($j = 1$) ou second échantillon ($j = 2$). Pour fixer les idées, on admettra que la sommation est étendue aux éléments du premier échantillon. Les valeurs de la fonction de rang $\varphi(R_{i,n})$ s'appellent *eodes*. La répartition des statistiques de ces tests converge rapidement vers la répartition normale lorsque la taille des échantillons croît.

Les tests proposés plus bas sont convergents pour l'hypothèse d'hétérogénéité lorsque cette hétérogénéité est due à l'inégalité des paramètres de localisation (par exemple, les moyennes, les médianes) des répartitions. Si l'on a affaire à deux échantillons, on peut écrire les hypothèses concurrentes sous la forme :

$$\left\{\begin{array}{ll} H_{11}: F_1(x) = F_2(x-\mu), \ \mu \neq 0: & \text{les répartitions sont décalées l'une par rapport à l'autre;} \\ H_{12}: F_1(x) = F_2(x-\mu), \ \mu > 0: & \text{la deuxième répartition est déplacée à gauche par rapport à la première;} \\ H_{13}: F_1(x) = F_2(x-\mu), \ \mu < 0: & \text{la deuxième répartition est déplacée à droite par rapport à la première,} \end{array}\right. \tag{11.25}$$

l'hypothèse nulle (c'est-à-dire l'égalité des fonctions de répartition) étant H_0: $\mu_0 = 0$. Ces tests peuvent aussi être utilisés pour les hypothèses suivantes:

$$H_{21} : F_1(x) < F_2(x);$$
$$H_{22} : F_1(x) > F_2(x), \tag{11.26}$$

autrement dit, la fonction de répartition du deuxième échantillon est « stochastiquement strictement inférieure » et « stochastiquement strictement supérieure » à celle du premier respectivement pour H_{21} et H_{22}.

Test de Wilcoxon-Mann-Whitney. La statistique de ce test est de la forme

$$S = \sum_{i=1}^{n_1} R_{i,n}. \tag{11.27}$$

Les codes sont ici les rangs des observations. La statistique S est parfois appelée statistique de la somme des rangs. On fait souvent appel à la statistique équivalente

$$U = S - \frac{1}{2} n_1 (n_1 + 1). \tag{11.28}$$

Si l'hypothèse nulle est vraie, on a

$$\mathbf{E}S = \frac{1}{2} n_1 (n_1 + n_2 + 1); \quad \mathbf{E}U = \frac{1}{2} n_1 n_2; \tag{11.29}$$

$$\mathbf{V}S = \mathbf{V}U = \frac{1}{12} (n_1 + n_2) n_1 n_2. \tag{11.30}$$

Si n_1, $n_2 \to \infty$ de telle sorte que $0 < \lim n_1/n_2 = \lambda < \infty$, la fonction de répartition de U converge vers la fonction de répartition d'une loi normale dont la moyenne et la variance se déduisent des relations (11.29) et (11.30) respectivement. La convergence vers la loi normale est très rapide et elle est déjà efficace si n_1 et $n_2 > 8$ [77].

L'approximation suivante *)

$$P(S \leqslant t) \approx \Phi(x) + \varphi(x)(x^3 - 3x) \frac{n_1^2 + n_2^2 + n_1 n_2 + n_1 + n_2}{20 n_1 n_2 (n_1 + n_2 + 1)}, \tag{11.31}$$

où $\Phi(x)$ et $\varphi(x)$ sont respectivement la fonction de répartition et la densité de probabilité de la loi normale réduite (cf. n° 12.1.2) de la variable

$$x = (t - \mathbf{E}S + 0{,}5)/\sqrt{\mathbf{V}S}$$

est encore plus exacte.

*) Fix E., Hodges J. L. *Significance probabilites of the Wilcoxon test.* — AMS, 1955, vol. 26, 301-312.

Test des codes normaux (Fisher-Yates-Terry-Hoeffding)*). La statistique de ce test s'écrit

$$C = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathrm{E}\,(R_{i,\,n},\ n), \tag{11.32}$$

où E (r, n) est l'espérance mathématique de la statistique de rang r (cf. n° 12.2.10) dans un échantillon de taille $n = n_1 + n_2$ extrait d'une population générale suivant la loi normale réduite.

Si l'hypothèse nulle est vraie, alors

$$\mathrm{E}C = 0; \tag{11.33}$$

$$\mathrm{V}C = \frac{n_2}{n\,(n-1)\,n_1} \sum_{i=1}^{n} [\mathrm{E}\,(r,\ n)]^2. \tag{11.34}$$

La répartition limite converge vers la répartition normale lorsque $n_1,\ n_2 \to \infty$; la quantité $\left(\frac{1}{n} \sum_{r=1}^{n} [\mathrm{E}\,(r,\ n)]^2\right)^{1/2} = 0{,}97$ pour $n = 50$, elle tend ensuite vers 1, de sorte que $\mathrm{V}C \approx n_2/(n \cdot n_1)$ pour de grands n. On se sert parfois du test de Van der Waerden [138] de statistique

$$V = \frac{1}{n_1} \sum_{i=1}^{n_1} \psi\left(\frac{R_{i,\,n}}{n+1}\right), \tag{11.35}$$

où $\psi\,(\cdot)$ est la fonction inverse de la fonction de répartition de la loi normale réduite (cf. n° 12.2.1). Ce test est asymptotiquement équivalent au test C. Si l'hypothèse nulle est vraie, alors

$$\mathrm{E}V = 0;$$

$$\mathrm{V}V = \frac{n_2}{n\,(n-1)\,n_1} \sum_{i=1}^{n} \psi^2\left(\frac{r}{n+1}\right).$$

L'application des tests S, C et V implique la procédure de calcul suivante.

1. Les échantillons sont réunis et l'échantillon global est rangé par ordre de grandeur croissante des observations.

2. On calcule la valeur de la statistique du test K à l'aide des rangs du premier (ou du deuxième) échantillon (ce peut être une

*) Terry M. E. *Some rank order tests which are most powerful against, specific alternatives.* — AMS, 1952, vol. 23, 346-366.

valeur de S, C ou V), puis, la quantité $A_K = (K - \mathrm{E}K)/\sqrt{\mathrm{V}K}$.

3. On compare la valeur de A_K avec les quantiles de la loi normale réduite (ou bien on calcule la valeur de la fonction de répartition $\Phi(A_K)$ de la loi normale réduite). Selon l'hypothèse concurrente, la région de rejet de l'hypothèse nulle au seuil de α est définie par les inégalités (sous réserve que le calcul de la statistique du test ait été effectué sur les éléments du premier échantillon):

$$|A_K| \geqslant u_{1-\frac{\alpha}{2}} \text{ pour } H_{11}';$$

$$A_K \geqslant u_{1-\alpha} \text{ pour } H_{12} \text{ et } H_{21};$$

$$A_K < u_\alpha \text{ pour } H_{13} \text{ et } H_{22}.$$

Si les échantillons sont de petite taille, on peut utiliser les tables des valeurs critiques des statistiques respectives *) pour obtenir un résultat plus exact.

De tous les tests de rang concernant les hypothèses d'égalité des paramètres de localisation pour n assez grand, le plus puissant est le test des codes normaux, et le moins puissant, le test de Wilcoxon (si n n'est pas grand, le test de Wilcoxon peut être plus puissant que celui des codes normaux pour certains types de fonctions de répartition modèles). En particulier, pour les fonctions de répartition normales, le test des codes normaux est de même puissance que le test t envisagé au n° 11.2.7. Une telle analyse comparative des propriétés des tests est accessible dans [77].

11.2.4. Test non paramétrique de l'hypothèse d'égalité des variances).** L'hypothèse nulle est la même que la précédente:

$$H_0 : F_1(x) = F_2(x) = F(x).$$

Mais l'hypothèse concurrente porte maintenant sur l'inégalité des variances des répartitions

$$H_1 : F_1(x) = F_2(x/\tau) \quad \text{ou} \quad \sigma_1/\sigma_2 = \tau \ (\tau > 0).$$

Ceci étant, on peut dégager une hypothèse bilatérale:

H_{11} : $\tau \neq 1$, les variances des répartitions sont distinctes.

et les hypothèses unilatérales:

---

*) De telles tables sont disponibles pour les tests de Wilcoxon et de Van der Waerden dans [27]; une table des valeurs critiques pour le test des codes normaux pour $n_1 + n_2 \leqslant 20$ est accessible dans l'article: Klotz J. *On the normal scores two-sample rank test*. JASA, vol. 59, 1964, 652-664.

**) En toute rigueur, on aurait dû mentionner la différence des paramètres d'échelle. Mais on admet que les fonctions de répartition $F_1(x)$ et $F_2(x)$ possèdent une variance et l'hypothèse d'égalité des paramètres d'échelle est équivalente à l'hypothèse d'égalité des variances.

H_{12} : $\tau > 1$, la variance de la première répartition est $>$ à celle de la deuxième;

H_{13} : $\tau < 1$, la variance de la première répartition est $<$ à celle de la deuxième.

L'hypothèse nulle peut être mise sous la forme:

H_0 : $\tau = 1$, les variances des deux répartitions sont égales.

Pour que l'on puisse appliquer les tests de rang du paramètre d'échelle, il faut que les deux répartitions soient de même type, que leurs paramètres de localisation prennent la même valeur et que leurs densités de probabilité soient continues.

Le test d'égalité des variances cité plus bas appartient à la classe des tests de rang linéaires de la forme (11.24).

Test de Klotz *). La statistique de ce test est

$$V_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\psi\left(\frac{R_{i,n}}{n+1}\right)\right]^2.$$

Donc, les codes de ce test sont les carrés de ceux du test de Van der Waerden. L'espérance mathématique et la variance de cette statistique s'écrivent:

$$\mathrm{E}V_1 = \frac{n_1}{n} \sum_{r=1}^{n} \left[\psi\left(\frac{r}{n+1}\right)\right]^2;$$

$$\mathrm{V}V_1 = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{n} \left[\psi\left(\frac{r}{n+1}\right)\right]^4 - \frac{n_2}{n_1(n-1)} (\mathrm{E}V_1)^2.$$

La statistique V_1 suit une loi asymptotiquement normale de moyenne $\mathrm{E}V_1$ et de variance $\mathrm{V}V_1$. Pour les petites valeurs de n_1 et n_2 ($n_1 + n_2 \leqslant 20$), il existe une table de points critiques de la fonction de répartition exacte de la statistique V_1 dans le cas de l'hypothèse nulle. Pour les grandes valeurs de n_1 et n_2, on peut se servir d'une fonction de répartition asymptotique. La méthode d'utilisation de ce test dans le cas des hypothèses concurrentes bilatérale et unilatérales, est la même que celle appliquée pour vérifier des hypothèses de même nature par les tests de rang du paramètre de localisation.

Il existe d'autres tests de rang linéaires [67] pour vérifier l'égalité des variances, par exemple, le test de Mood qui est identique au test de Wilcoxon de statistique

$$S_1 = \sum_{i=1}^{n_1} \left(R_{i,n} - \frac{1}{2}(n+1)\right)^2.$$

*) Klotz J. *Nonparametric tests for scale.* — AMS, 1962, vol. 33, p. 498-512. Cet article contient la table des valeurs critiques pour $n_1 + n_2 \leqslant 20$.

Mais le test de Klotz est en général plus puissant que le test S_1.

La condition d'égalité des paramètres de localisation (la condition $\mu = 0$) est essentielle à l'application du test de Klotz (de même que les autres tests de rang du paramètre d'échelle), sinon le seuil de signification de la statistique du test risque d'être fortement altéré. Si donc l'on présume que ces paramètres sont inégaux, il faut les estimer préalablement pour chaque échantillon à l'aide, par exemple, des médianes empiriques ou d'autres estimations stables du paramètre de localisation (cf. chapitre 10), retrancher les résultats obtenus des éléments de l'échantillon correspondant et appliquer ensuite le test V_1.

Si les fonctions de répartition $F_1(x)$ et $F_2(x)$ sont normales, le test de Klotz possède pratiquement la même puissance que le meilleur test de Fisher pour ce cas (cf. n° 11.2.8).

11.2.5. Tests de rang pour $k > 2$. Les tests de rang d'homogénéité se généralisent de la manière suivante pour $k > 2$. On ordonne l'échantillon global de taille $n_1 + n_2 + \ldots + n_k = n$ et on affecte à chaque observation un rang correspondant à sa position dans cet échantillon. Pour chacune des k répartitions on calcule la quantité

$$K_i = \sum_{j=1}^{n_i} \varphi(R_{j,n}), \quad i = 1, \ldots, k,$$

où la sommation est étendue à toutes les observations appartenant au i-ième échantillon. Chaque quantité K_i suit une loi asymptotiquement normale. La somme $S_0 = \sum_{i=1}^{n} K_i$ est connue, c'est-à-dire que les quantités K_i sont reliées par une seule relation linéaire, donc la variable aléatoire

$$H = \sum_{i=1}^{k} \frac{(K_i - \mathrm{E}K_i)^2}{\mathrm{V}K_i}$$

suit une loi du χ^2 à $k - 1$ degrés de liberté si l'hypothèse nulle est vraie. En particulier, si pour K_i on prend la somme des rangs comme dans le test de Wilcoxon, on débouche sur le test de Kruskal-Wallis [67] de statistique

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{S_i^2}{n_i} - 3(n+1), \tag{11.36}$$

où S_i est la somme des rangs des observations du i-ième échantillon de l'échantillon global.

11.2.6. Test de symétrie des répartitions. Le test de symétrie des répartitions est un problème qui se pose dans l'analyse des restes des modèles de régression, en analyse de variance et dans l'estimation stable. Considérons le test de symétrie par rapport à un point fixe μ. On suppose que la loi de répartition admet une densité de probabilité continue $f(x)$. L'hypothèse de symétrie peut alors être mise sous la forme:

$$H_0 : f(\mu + x) = f(x - \mu) \tag{11.37}$$

ou en termes de fonctions de répartition

$$H_0 : F(\mu + x) = 1 - F(x - \mu). \tag{11.38}$$

L'hypothèse (11.37) affirme que la densité de probabilité $f(x)$ est symétrique et que le centre de symétrie est le point μ.

Les tests *non paramétriques de l'hypothèse de symétrie* sont basés sur les *rangs absolus* (par rapport au point μ).

Soit $x_1, \ldots, x_n$ un échantillon pour lequel on teste l'hypothèse (11.37), (11.38). Introduisons les observations transformées

$$z_i = | x_i - \mu | \tag{11.39}$$

et formons la série variationnelle $z_{(1)}, \ldots, z_{(n)}$. On appellera le rang de la quantité z_i *rang absolue* de x_i (par rapport au point μ) et on le désignera par $R^+_{i,n}$. Cette transformation ramène le test de l'hypothèse de symétrie au test d'homogénéité de deux répartitions formées respectivement avec les ailes gauche et droite (par rapport à μ) de la répartition primitive.

Soit I^+ l'ensemble des indices des observations x_i pour lesquelles $x_i - \mu > 0$, autrement dit, si $i \in I^+$, alors $x_i - \mu > 0$. Les tests envisagés ici sont identiques aux tests de rang d'homogénéité du n° 11.2.3.

Le test de Wilcoxon pour un échantillon *) utilise la statistique

$$S^+ = \sum_{i \in I^+} R^+_{i,n}. \tag{11.40}$$

Si l'hypothèse nulle est vraie, on a [67]

$$\begin{cases} \mathrm{E}S^+ = \frac{1}{4} n(n+1); \\ \mathrm{V}S^+ = \frac{1}{24} n(n+1)(2n+1). \end{cases} \tag{11.41}$$

*) Ce test a été introduit dans l'article: Wilcoxon F. *Individual comparaisons by ranking methods.* — Biometrics Ball. 1945. 1, 80-83.

Le *test de Fraser-Klotz* *) (ou test des codes normaux) est basé sur la statistique

$$C^+ = \sum_{i \in I^+} \mathrm{E}^+ (R^+_{i,n}, n), \tag{11.42}$$

où $\mathrm{E}^+ (k, n) = \mathbf{E} (| V |_{(k)})$ est l'espérance mathématique de la statistique de rang k dans la série variationnelle de longueur n constituée par les valeurs *absolues* des variables aléatoires normales réduites $V_1, \ldots, V_n$.

Si l'hypothèse nulle est vraie, on a [67]

$$\begin{cases} \mathbf{E}C^+ = \dfrac{n}{\sqrt{2\pi}}\,; \\ \mathbf{V}C^+ = \dfrac{1}{4} \displaystyle\sum_{r=1}^{n} [\mathrm{E}^+ (r, n)]^2. \end{cases} \tag{11.43}$$

L'*analogue du test de Van der Waerden* est asymptotiquement semblable à celui de Fraser-Klotz. La statistique de ce test est de la forme [67]

$$V^+ = \sum_{i \in I} \psi \left(\frac{1}{2} + \frac{1}{2} \frac{R^+_{i,n}}{n+1} \right) \tag{11.44}$$

L'espérance mathématique et la variance s'écrivent

$$\begin{cases} \mathbf{E}V^+ = \dfrac{1}{2} \displaystyle\sum_{r=1}^{n} \psi \left(\frac{1}{2} + \frac{1}{2} \frac{r}{n+1} \right); \\ \mathbf{V}V^+ = \dfrac{1}{4} \displaystyle\sum_{r=1}^{n} \left[\psi \left(\frac{1}{2} + \frac{1}{2} \frac{r}{n+1} \right) \right]^2 \end{cases} \tag{11.45}$$

où ψ ( ) est la fonction inverse de la fonction de répartition de la loi normale réduite.

Tous les tests de rang introduits suivent une loi asymptotiquement normale dont les paramètres sont définis par les formules (11.41), (11.43) et (11.45) respectivement. L'application de ces tests se ramène à la réalisation de la procédure suivante.

1. A l'échantillon de départ $x_1, \ldots, x_n$ on associe l'échantillon $z_i = | x_i - \mu |$ $(i = 1, \ldots, n)$.

---

*) Ce test a été introduit dans l'article : Fraser D.A.S. *Most powerful rank type tests.* — Ann. Math. Stat. 1957, vol. 28, 1040-1043. Les tables des codes et des valeurs critiques pour $n \leqslant 10$ figurent dans l'article : Klotz J. *Small sample power and efficiency for the one sample Wilcoxon and normal scores tests.* — Ann. Math. Stat. 1963, 34, 624—632.

2. On range les quantités z_i par ordre de grandeur croissante:

$$z_{(1)} \leqslant z_{(2)} \leqslant \ldots \leqslant z_{(n)}.$$

3. On détermine les rangs des observations $z_{(1)}, \ldots, z_{(n)}$ correspondant aux observations primitives pour lesquelles $x_i - \mu \geqslant 0$ (ou $x_i - \mu < 0$).

4. On calcule la statistique du test K^+ à l'aide des formules (11.40), (11.42) ou (11.44).

5. On calcule la quantité $A^+ = (K^+ - \mathbf{E}K^+)/\sqrt{\mathbf{V}K^+}$.

6. On rejette l'hypothèse de symétrie si $|A^+|$ est trop grand, plus précisément, si est réalisée l'une des inégalités suivantes: $A^+ > u_{1-\alpha/2}$ ou $A^+ < u_{\alpha/2}$, où α est le seuil de signification donné de l'hypothèse nulle. Donc, la région critique des critères de symétrie (11.40), (11.42), (11.44) est bilatérale.

Le centre de symétrie théorique est souvent inconnu. Dans ces cas, pour tester les hypothèses (11.37), (11.38), on prend pour point μ une quelconque estimation du paramètre de localisation, par exemple, la moyenne arithmétique, la médiane ou une estimation stable de ce paramètre (cf. chapitre 10). Dans ces conditions, les tests non paramétriques envisagés plus haut sont approximatifs.

11.2.7. Traitement des coïncidences. L'application des tests de rang implique la continuité des lois de répartitions des variables aléatoires observées. Mais, dans la pratique, on est toujours confronté à des valeurs discrètes de la variable aléatoire soit que cette variable est de nature discrète, soit que les valeurs observées sont arrondies ou groupées. Ce qui fait que certains groupes d'observations sont confondus. Voyons quelques méthodes d'application des tests de rang dans le cas de coïncidences.

Si les observations confondues appartiennent à un même échantillon, aucun problème ne se pose: pour rangs on peut prendre les numéros de ces observations dans un ordre quelconque. Si le groupe de coïncidence contient des éléments des deux échantillons, les méthodes de traitement les plus couramment utilisées sont les suivantes [67].

Méthode du rang aléatoire. Aux observations confondues on attribue de façon aléatoire (c'est-à-dire avec la même probabilité) des numéros (rangs). Dans ce cas, la théorie de la répartition de la statistique du test pour l'hypothèse nulle reste en vigueur et on peut se servir des habituelles tables et répartitions limites. Mais la puissance du test sera plus petite que par la méthode du code moyen.

Méthode du code moyen. A toutes les observations du premier (ou du deuxième) échantillon, contenues dans le groupe de coïncidence, on attribue la valeur moyenne du code pour les observations de ce groupe. Dans ce cas, la loi limite de la statistique du test restera normale. L'espérance mathématique ne varie pas, quant à la va-

riance, elle diminue. Compte tenu des coïncidences, les variances des statistiques des tests de Wilcoxon, des codes normaux et de Van der Waerden sont données par les formules

$$\begin{cases} \widetilde{\mathrm{V}}S = \mathrm{V}S\left(1 - \dfrac{\sum\limits_{j=1}^{g} \tau_j(\tau_j^2-1)}{n^3-n}\right); \\ \widetilde{\mathrm{V}}C = \dfrac{n_1 n_2}{n(n-1)} \sum\limits_{j=1}^{g} \tau_j \overline{\mathrm{E}}_j^2; \\ \widetilde{\mathrm{V}}V = \dfrac{n_1 n_2}{n(n-1)} \sum\limits_{j=1}^{g} \tau_j \overline{\psi}_j^2, \end{cases} \tag{11.46}$$

où g est le nombre de groupes d'observations; τ_j ($j = 1, \ldots, g$) le nombre d'observations du groupe j ($\sum\limits_{j=1}^{g} \tau_j = n$); $\overline{\mathrm{E}}_j$ et $\overline{\psi}_j$ les codes moyens sur le groupe j respectivement pour les tests des codes normaux et de Van der Waerden

Ces valeurs des variances doivent être utilisées dans les tests de rang respectifs. En vertu de (11.46), la statistique du test de Kruskal-Wallis (11.36) se modifie de la manière suivante dans le cas de coïncidences:

$$H' = H \Big/ \left(1 - \sum_{j=1}^{g} \frac{\tau_j(\tau_j^2-1)}{n^3-n}\right). \tag{11.47}$$

Pour le *test de symétrie*, considérons seulement le cas où la répartition admet une densité continue au centre hypothétique μ de symétrie, autrement dit, la probabilité d'obtenir une observation égale à μ est nulle. On a alors les formules suivantes pour les variances:

$$\mathrm{V}S^+ = \frac{1}{4} \sum_{j=1}^{g} \left[\frac{\tau_j(\tau_j+1)}{2}\right]^2; \tag{11.48}$$

$$\mathrm{V}C^+ = \frac{1}{4} \sum \tau_j [\overline{\mathrm{E}}_j^+]^2; \tag{11.48'}$$

$$\mathrm{V}V^+ = \frac{1}{4} \sum \tau_j [\overline{\psi}_j^+]^2, \tag{11.48''}$$

où $\overline{\mathrm{E}}_j^+$ est la valeur moyenne des codes $\mathrm{E} \mid x \mid_{(i)}$ pour les observations du groupe j. On détermine de façon analogue les quantités $\overline{\psi}_j^+$ pour le test de Van der Waerden.

11.2.8. Test d'homogénéité des populations normales (cas scalaire). A strictement parler, les tests décrits plus bas (le test de Student, de Fisher du rapport dispersionnel, de Bartlett, etc.) ne s'appliquent qu'aux échantillons (9.3) extraits d'une population *normale*: dans ce cas, on conçoit sans peine que le résultat positif du test simultané de l'homogénéité des valeurs moyennes (c'est-à-dire de l'hypothèse (9.3b)) et des variances (c'est-à-dire de l'hypothèse (9.3c)) suffit pour conclure que l'hypothèse d'homogénéité des fonctions de répartition respectives (c'est-à-dire l'hypothèse (9.3a)) est vraie. Des analyses spéciales ont montré toutefois que le test de Student est le plus *stable* pour les écarts des populations générales étudiées par rapport aux normales. Ce qui signifie qu'*il peut être appliqué aux échantillons extraits de populations générales non normales*, à la seule condition que les véritables valeurs du seuil de signification et de la puissance du test diffèrent peu des valeurs données.

Le *test d'homogénéité de deux échantillons* ($k = 2$) est basé sur la statistique critique

$$\hat{t}(n_1 + n_2 - 2) = \frac{\bar{x}_1(n_1) - \bar{x}_2(n_2)}{\tilde{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

où $\bar{x}_1(n_1)$ et $\bar{x}_2(n_2)$ sont les valeurs arithmétiques moyennes des observations respectivement du premier et du deuxième échantillon, $\tilde{s}^2$ l'estimation globale de la variance sur l'ensemble des deux échantillons, c'est-à-dire que

$$\tilde{s}^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\bar{s}_1^2(n_1) + (n_2 - 1)\bar{s}_2^2(n_2)\right]; \qquad (11.49)$$

$$\bar{s}_j^2(n) = \frac{1}{n_j - 1}\sum_{i=1}^{n}(x_i - \bar{x}_j(n))^2, \quad j = 1,\ 2. \qquad (11.50)$$

On sait (cf. n° 6.2.2) que si l'hypothèse d'homogénéité (9.3a) est vraie, la statistique $\hat{t}(n_1 + n_2 - 2)$ suit la loi de Student à $n_1 + n_2 - 2$ degrés de liberté. Donc, pour vérifier l'hypothèse d'homogénéité, il faut trouver dans les tables des points de pourcentage de la loi de Student (cf. par exemple [27], table 3.2) le point $t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$ de pourcentage $100\,\frac{\alpha}{2}$ de la loi de Student à $n_1 + n_2 - 2$ degrés de liberté. Si

$$|\hat{t}(n_1 + n_2 - 2)| \leqslant t_{\frac{\alpha}{2}}(n_1 + n_2 - 2), \qquad (11.51)$$

l'hypothèse d'homogénéité des échantillons $\{x_{11}, x_{12}, \ldots, x_{1n_1}\}$ et $\{x_{21}, x_{22}, \ldots, x_{2n_2}\}$ est acceptée (sinon elle est refusée).

Remarque 1. Une « trop grande » valeur de $\hat{t}(n_1 + n_2 - 2)$, c'est-à-dire une valeur violant l'inégalité (11.51), peut être le résultat d'une importante différence aussi bien entre les moyennes (c'est-à-dire la non-réalisation de l'hypothèse (9.3b)) qu'entre les variances (c'est-à-dire la non-réalisation de l'hypothèse (9.3c)).

Si l'on veut mettre en évidence la cause de l'hétérogénéité des échantillons, il faut vérifier accessoirement l'homogénéité des variances, c'est-à-dire l'hypothèse (9.3c) pour $k = 2$ (le test de l'homogénéité des variances constitue parfois le seul objectif de l'analyse). Le test d'homogénéité de deux variances empiriques est basé sur la statistique critique

$$\hat{F}(n_1-1,\ n_2-1) = \frac{\dfrac{1}{n_1-1}\sum\limits_{i=1}^{n_1}(x_{1i}-\bar{x}_1(n_1))^2}{\dfrac{1}{n_2-1}\sum\limits_{j=1}^{n_2}(x_{2j}-\bar{x}_2(n_2))^2},$$

qui, on le sait (cf. n° 6.2.3), suit, dans le cas où l'hypothèse (9.3a) est vraie, une loi $F(n_1 - 1,\ n_2 - 1)$ de Fisher à $(n_1 - 1,\ n_2 - 1)$ degrés de liberté. Si donc

$$F_{1-\frac{\alpha}{2}}(n_1-1,\ n_2-1) \leqslant \hat{F}(n_1-1,\ n_2-1) \leqslant$$
$$\leqslant F_{\frac{\alpha}{2}}(n_1-1,\ n_2-1), \qquad (11.52)$$

l'hypothèse d'homogénéité des variances est acceptée (sinon, refusée). Dans les inégalités (11.52), $F_q(m_1, m_2)$ désigne le point de pourcentage $100q$ d'une loi F à (m_1, m_2) degrés de liberté (voir, par exemple, la table 3.5c [27]), α le seuil de signification donné du test.

Enfin, il est possible que les variances σ_1^2 et σ_2^2 soient différentes (c'est-à-dire que l'une au moins des inégalités (11.52) est violée), alors que l'on s'intéresse à l'homogénéité des valeurs moyennes, c'est-à-dire à l'hypothèse (9.3b). Dans ce cas, on peut construire un test *approché* basé sur la statistique critique

$$\tilde{t}(n_1,\ n_2,\ c) = \frac{\bar{x}_1(n_1) - \bar{x}_2(n_2)}{\sqrt{\dfrac{1}{n_1}\bar{s}_1^2(n_1) + \dfrac{1}{n_2}\bar{s}_2^2(n_2)}},$$

où $\bar{s}_j^2(n_j)$ sont définis par la formule (11.50) et

$$c = \frac{\dfrac{1}{n_1}\bar{s}_1^2(n_1)}{\dfrac{1}{n_1}\bar{s}_1^2(n_1) + \dfrac{1}{n_2}\bar{s}_2^2(n_2)}.$$

La dépendance de la statistique critique t par rapport au paramètre auxiliaire c est prise en considération dans les tables des points $\tilde{t}_{\frac{\alpha}{2}}(n_1, n_2, c)$ de pourcentage de la loi de cette variable aléatoire. Ces points de pourcentage peuvent être trouvés, par exemple, dans la table 4.4 de [27] (dans ces tables $\tilde{t}_{\frac{\alpha}{2}}(n_1, n_2, c)$ est représenté par $V(c, \nu_1, \nu_2, Q)$, où $\nu_k = n_k - 1$ et $Q = \frac{\alpha}{2}$).

Si, par exemple,

$$|\tilde{t}(n_1, n_2, c)| \leqslant \tilde{t}_{\frac{\alpha}{2}}(n_1, n_2, c),$$

on en déduit que les moyennes vérifiées sont homogènes.

Remarque 2. Dans tous les cas où la différence entre les moyennes n'est vérifiée que dans un sens (par exemple, on vérifie s'il « est possible de considérer que la moyenne du premier échantillon est significativement supérieure à celle du deuxième échantillon »), pour ordre du point de pourcentage il faut prendre α et non pas $\frac{\alpha}{2}$.

Le *test d'homogénéité de plusieurs échantillons* ($k > 2$, cf. (9.3)) est basé sur la statistique critique

$$\hat{F}^{(k)}(k-1, n_1+n_2+\ldots+n_k-k) = \frac{\frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{x}_k(n_k)-\bar{x}(n))^2}{\tilde{s}^2(n)}, \tag{11.53}$$

où

$$\bar{x}(n) = \frac{1}{n_1+\ldots+n_k}\sum_{j=1}^{k} n_j\bar{x}_j(n_k);$$
$$\tilde{s}^2(n) = \frac{1}{n_1+\ldots+n_k-k}\sum_{j=1}^{k}(n_j-1)\,\bar{s}^2(n_j), \tag{11.54}$$

où $\bar{x}(n)$ est la moyenne arithmétique générale calculée sur la réunion des k échantillons, et les variances $\bar{s}_j^2(n_j)$ empiriques sont calculées à l'aide de la formule (11.50).

On démontre que si l'hypothèse (9.3a) est vraie, la statistique (11.53) suit une loi F à $(k-1, n_1+n_2+\ldots+n_k-k)$ degrés

de liberté (cf. n° 6.2.3). Si donc

$$\hat{F}^{(k)}(k-1,\ n_1+n_2+\ldots+n_k-k) \leqslant F_\alpha(k-1,\ n_1+\ldots+n_k-k), \qquad (11.55)$$

on accepte l'hypothèse d'homogénéité des échantillons (9.3).

Remarque 3. De même que dans le cas de deux échantillons, une « trop grande » valeur de la statistique critique (11.53), c'est-à-dire une valeur pour laquelle est violée l'inégalité (11.55), peut être la conséquence d'une différence significative aussi bien des moyennes (c'est-à-dire de la non-réalisation de l'hypothèse (9.3b)) que des variances (c'est-à-dire de la non-réalisation de l'hypothèse (9.3c)). On aura donc intérêt à vérifier l'homogénéité d'une série (de plus de deux) variances empiriques. Cette vérification peut être effectuée, par exemple, avec le test de Bartlett ([77], n° 24.9) basé sur la statistique critique

$$\lambda = q \sum_{i=1}^{k} (n_i - 1) \ln \left(\frac{\tilde{s}^2(n)}{\bar{s}_i^2(n_i)} \right), \qquad (11.56)$$

où $\tilde{s}^2(n)$ et $\bar{s}_i^2(n_i)$ sont définies respectivement par (11.54) et (11.50), et

$$q = \left[1 + \frac{1}{3(k-1)} \cdot \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n_1 + \ldots + n_k - k} \right) \right]^{-1}. \qquad (11.57)$$

M. S. Bartlett a prouvé *) que si $\min(n_1, n_2, \ldots, n_k) > 3$ et que l'hypothèse d'égalité des variances (cf. 9.3c) soit vraie, la statistique λ suit approximativement une loi du χ^2 à $k-1$ degrés de liberté. Donc, si $\lambda < \chi^2_\alpha(k-1)$, l'hypothèse d'homogénéité des variances empiriques est acceptée, sinon, elle est rejetée.

11.2.9. Test d'homogénéité de populations normales multidimensionnelles. Dans le cas multidimensionnel, l'hypothèse d'homogénéité se formule comme en dimension un (cf. (11.14)):

$$H_0 : F_1(X) = F_2(X) = \ldots = F_k(X) = F(X), \qquad (11.58)$$

sauf que X est à p dimensions.

Le choix des méthodes de test de l'hypothèse (11.58) est bien plus restreint qu'en dimension un. En fait, on ne peut appliquer que des tests paramétriques en postulant que chaque fonction de répartition $F_i(X)$ $(i = 1, \ldots, k)$ est la fonction de répartition d'une

*) Bartllet M.S. *Properties of sufficiency of statistical tests.* — Proc. Roy. Soc., 1937, A-160, p. 268-282.

loi normale multidimensionnelle. Comme en dimension un, certains de ces tests, par exemple, les tests du Λ, de T^2, sont stables pour les désaccords par rapport à la fonction de répartition normale et, par suite, peuvent être appliqués aux échantillons non normaux.

On sait (cf. § 6.1) qu'une fonction de répartition normale multidimensionnelle est entièrement définie par le vecteur des moyennes M_i et la matrice des covariances Σ_i ($i = 1, \ldots, k$). De façon respective, les statistiques des tests envisagés plus bas sont des fonctionnelles des estimations empiriques des paramètres $\hat{M}_i$ et $\overline{\Sigma}_i$ ($i = 1, \ldots, k$).

Dans le cas de deux échantillons ($k = 2$), la statistique critique du test de l'hypothèse d'homogénéité est la quantité [16].

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\hat{M}_1 - \hat{M}_2)' \, S^{-1} (\hat{M}_1 - \hat{M}_2). \tag{11.59}$$

Ceci étant, on admet *a priori* que les échantillons sont extraits de populations générales ayant une même matrice des covariances, c'est-à-dire que $\Sigma_1 = \Sigma_2$ et

$$S = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) S_1 + (n_2 - 1) S_2]. \tag{11.60}$$

Si l'hypothèse nulle est vraie, la variable $\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2$ suit une loi F à $(p, n_1 + n_2 - p - 1)$ degrés de liberté.

Dans le cas de plusieurs échantillons ($k > 2$), le test d'homogénéité (sachant que les matrices des covariances sont égales) est basé sur la Λ-statistique de [114]

$$\Lambda = \frac{\det(S)}{\det(C)},$$

où $S = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) S_i$ est la matrice des covariances de la population globale, $n = \sum_{i=1}^{k} n_i$;

$$C = S + \frac{1}{n} \sum_{i=1}^{k} n_i (\hat{M}_i - \hat{M}) (\hat{M}_i - \hat{M})'$$

la matrice des covariances estimée sur l'échantillon global obtenu en regroupant les k échantillons, et $\hat{M}$ le vecteur des moyennes de l'échantillon global.

En termes d'analyse de variance, S s'appelle matrice de *variance intragroupe*, $W = \frac{1}{n} \Sigma \, n_i (\hat{M}_i - \hat{M})(\hat{M}_i - \hat{M})'$, matrice de *variance intergroupe* et $C = S + W$, matrice de *variance totale*. On dé-

montre que pour $k = 2$ la quantité Λ est égale à

$$\Lambda = \frac{1}{(1+2T^2)},$$

c'est-à-dire que Λ est une fonction monotone de la statistique T et, par suite, pour $k = 2$ les tests de Λ et de T^2 conduisent aux mêmes résultats.

Les valeurs de Λ sont comprises dans l'intervalle [0, 1], et si l'hypothèse nulle est vraie, la valeur prise par Λ doit être proche de 1. Une valeur de Λ de beaucoup inférieure à 1 est une preuve d'hétérogénéité, c'est-à-dire que l'hypothèse nulle doit être rejetée. Mais la répartition exacte de Λ est très compliquée et on se sert dans la pratique de statistiques qui sont des fonctions de Λ.

Une de ces statistiques est la statistique $B(\Lambda)$ proposée par Bartlett [114] :

$$B(\Lambda) = -\left(n - \frac{p+k+1}{2}\right) \ln \Lambda. \tag{11.61}$$

Si l'hypothèse nulle est vraie, la loi de $B(\Lambda)$ se laisse approcher par une loi du χ^2 à p degrés de liberté.

Une autre statistique qu'il faut utiliser de préférence pour les échantillons de petite taille est la statistique proposée par Rao [114] :

$$U(\Lambda) = \frac{1-\Lambda^{1/n_\Lambda}}{\Lambda^{1/n_\Lambda}} \; \frac{\tilde{n} \cdot n_\Lambda + 1 - pk/2}{pk}. \tag{11.62}$$

Si l'hypothèse nulle est vraie, la loi de cette statistique se laisse approcher par une loi F à $(kp,\ \tilde{n}n_\Lambda + 1 - pk/2)$ degrés de liberté. Les quantités $\tilde{n}$ et n_Λ ne sont pas nécessairement entières et se calculent avec les formules :

$$\tilde{n} = n - (p - k + 1)/2\,;$$

$$n_\Lambda = \begin{cases} \sqrt{\dfrac{k^2p^2+4}{k^2+p^2-5}} & \text{si} \quad p^2+k^2 > 5\,; \\ 1 & \text{si} \quad p^2+k^2 = 5. \end{cases}$$

Dans les deux cas, la région critique est le domaine des grandes valeurs. Comme en dimension un, de « trop grandes » valeurs des tests (11.59), (11.61), (11.62) résultent de la violation de la condition d'égalité des matrices des covariances, bien que les moyennes diffèrent de façon non significative. Pour vérifier l'hypothèse

$$H_1 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma$$

on fait appel à la statistique [16]

$$\lambda = \frac{1}{2}\rho\left((n-k)\ln(\det(S)) - \sum_{i=1}^{k}(n_i-1)\ln(\det(\hat{\Sigma}_i))\right), \qquad (11.63)$$

où

$$\rho = 1 - \frac{2p^2+3p-1}{6(k-1)(p+1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right).$$

Si H_1 est vraie, cette statistique suit une loi asymptotique du χ^2 à $\frac{1}{2}(k-1)p(p+3)$ degrés de liberté. La méthode d'application du test (11.63) est la même que pour le test d'égalité des variances (11.56) en dimension un. La statistique (11.63) est confondue avec la statistique (11.56) en dimension un.

11.3. Test d'indépendance et de stationnarité d'une série d'observations

Avant de soumettre les résultats des observations à un traitement statistique, il faut s'assurer qu'ils forment bien un échantillon aléatoire, qu'ils sont stochastiquement indépendants (c'est-à-dire que le choix des observations n'est pas subjectif, que les observations ne dépendent pas de leur rang, par exemple la moyenne de la répartition étudiée ne doit pas subir de déplacement de nature monotone ou cyclique à mesure que le rang des observations croît, etc.).

11.3.1. Test des blocs, basé sur la médiane de l'échantillon. Soit donné un échantillon $x_1, x_2, \ldots, x_n$ extrait d'une certaine population générale. Rangeons les éléments de cet échantillon dans l'ordre de grandeur croissante: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Pour valeur empirique de la médiane $\hat{x}_{\text{méd}}(n)$, on prend (cf. chap. 8) l'élément médian (par sa position) de la série variationnelle, c'est-à-dire que

$$\tilde{x}_{\text{méd}}(n) = x_{\left(\frac{n+1}{2}\right)} \quad \text{si } n \text{ est impair};$$

$$\hat{x}_{\text{méd}}(n) = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) \quad \text{si } n \text{ est pair.}$$

On revient ensuite à l'échantillon de départ $x_1, x_2, \ldots, x_n$ et à la place de chaque x_i on écrira « + » si $x_i > \hat{x}_{\text{méd}}(n)$ et « — » si $x_i < \hat{x}_{\text{méd}}(n)$ (les termes égaux à $\hat{x}_{\text{méd}}(n)$ sont omis dans cette suite de « + » et de « — »). La suite de « + » et de « — » obtenue se caractérise par le nombre total $\nu(n)$ de blocs et par la longueur $\tau(n)$ du plus

long bloc. Par bloc on entend une suite maximale de « + » ou de « — » (en particulier, un bloc peut être composé d'un seul « + » ou d'un seul « — », auquel cas sa longueur est égale à 1). Il est évident que si les observations sont stochastiquement indépendantes (c'est-à-dire que l'échantillon est aléatoire), l'alternance des « + » et des « — » doit être plus ou moins « aléatoire », autrement dit, cette suite ne doit pas contenir de trop longues suites de signes « + » ou « — », et par conséquent, le nombre total $\nu(n)$ de blocs ne doit pas être trop petit. En sorte que dans ce test, on a intérêt à considérer le couple de statistiques critiques $\{\nu(n);\ \tau(n)\}$.

Pour construire un test statistique *exact* d'indépendance, basé sur la statistique à deux dimensions $\{\nu(n);\ \tau(n)\}$, il aurait fallu préalablement trouver et tabuler la loi de probabilité à deux dimensions de cette statistique. Nous nous bornerons ici à exposer un test *approximatif*. Pour le construire nous utiliserons : une loi normale de paramètres $\left(\frac{n+2}{2},\ \frac{\sqrt{n-1}}{2}\right)$ approchant la loi (marginale) à une dimension de la variable aléatoire $\nu(n)$; la loi de Poisson (cf. n° 6.1.3) (de paramètre $\lambda = n\tau_0/2 + 1$) du nombre de blocs de longueur $\geqslant \tau_0$ (cf. [7], pages 297-298) ; et, enfin, la majoration de la probabilité

$$P\ \{\nu(n) > \nu_{0,95}(n),\ \tau(n) < \tau_{0,05}(n)\},$$

où $\nu_{0,95}(n)$ est le point de pourcentage 0,95 de la loi marginale de $\nu(n)$, et $\tau_{0,05}(n)$ le point de pourcentage 0,05 de la loi marginale de $\tau(n)$.

Ceci nous conduit en définitive à la règle suivante. Si l'une au moins des inégalités

$$\begin{aligned} &\nu(n) > \left[\frac{1}{2}\left(n+1-1{,}96\sqrt{n-1}\right)\right]; \\ &\tau(n) < [3{,}3 \log(n+1)] \end{aligned} \tag{11.64}$$

est violée, l'hypothèse d'indépendance stochastique des observations *est rejetée* à un seuil α compris entre 0,05 et 0,0975.

Exemple 11.5. Les tests à la durée de vie de 58 échantillons d'une chaîne de fabrication ont donné les résultats suivants : $\underline{38}$, $\underline{33}$, $\underline{29}$, $\underline{16}$, $\underline{44}$, $\underline{21}$, $\underline{16}$, $\underline{17}$, $\underline{19}$, 1, $\underline{22}$, $\underline{28}$, $\underline{22}$, 14, 7, 13, $\underline{21}$, 15, $\overline{34}$, $\overline{23}$, $\overline{15}$, $\overline{19}$, $\overline{32}$, $\overline{24}$, $\overline{14}$, $\overline{13}$, $\overline{22}$, $\overline{8}$, $\overline{30}$, $\overline{11}$, $\overline{15}$, $\overline{24}$, $\overline{26}$, 14, 11, $\overline{25}$, $\overline{17}$, $\overline{10}$, $\overline{19}$, 5, 6, $\overline{16}$, $\overline{7}$, $\overline{10}$, 1, 5, 2, $\overline{8}$, 14, $\overline{14}$, 15, $\overline{16}$, $\overline{13}$, 11, 9, 11, $\overline{19}$, $\overline{21}$. (Sont soulignées les valeurs empiriques à la place desquelles devraient figurer des « + » dans la suite correspondante.)

De nombreux facteurs qui conditionnent fortement la qualité des échantillons (la matière première, la compétence du personnel, etc.) sont soumis à d'inévitables oscillations dont la nature peut être aussi bien aléatoire que systématique. La question qui nous inté-

resse est de savoir si l'on a dûment tenu compte de cet état de choses en désignant la méthode de prélèvement des échantillons, autrement dit, les observations sont-elles stochastiquement indépendantes? Ainsi, le caractère de la variation des données empiriques dans le temps (les données empiriques se suivent dans l'ordre de prélèvement des échantillons) dénote une tendance systématique à la baisse de la durée de vie. Le test qui vient juste d'être décrit nous aidera à dissiper nos doutes.

Les calculs nécessaires nous donnent: $\hat{x}_{\text{méd}}(n) = 15{,}5$; $\tau(n) = 9$; $\nu(n) = 25$.

De sorte que des deux inégalités (11.64), seule la première est réalisée. Force est donc de reconnaître que le brassage des échantillons avant leur tirage n'a pas été très satisfaisant et, par conséquent, les résultats des observations présentés ci-dessus ne sont pas stochastiquement indépendants.

11.3.2. Test des séries « ascendantes » et « descendantes ». Ce test « décèle » tout déplacement progressif (au cours du sondage) de la valeur moyenne de la loi étudiée, qu'il soit monotone ou de nature plus générale, par exemple, périodique.

Comme dans le test précédent, on étudie une suite de signes « + » et « — » mais dont le principe de formation est différent. On part comme toujours d'une suite d'observations : un échantillon $x_1, x_2, \ldots, x_n$; on remplace le terme de rang i par « + » si $x_{i+1} > x_i$ et par « — » si $x_{i+1} < x_i$ (si deux ou plusieurs observations successives sont égales, on ne tient compte que de l'une d'elles). Il est évident qu'une suite de « + » correspondra à une série d'observations strictement croissantes (série ascendante) et une suite de « — », à une série d'observations strictement décroissantes (série descendante). Le test repose sur le même principe que le précédent : si l'échantillon est aléatoire (les observations sont indépendantes), le nombre total de blocs ne peut être trop petit, et leur longueur, pas trop grande.

En particulier, si le seuil de signification α est compris entre 0,050 et 0,0975, cette règle est exprimée par les inégalités:

$$\begin{aligned} &\nu(n) > \left[\frac{1}{3}(2n-1) - 1{,}96\sqrt{\frac{16n-29}{90}}\right]; \\ &\tau(n) < \tau_0(n), \end{aligned} \tag{11.65}$$

où $\nu(n)$ et $\tau(n)$ désignent comme toujours respectivement le nombre total de blocs et le nombre de signes « + » ou « — » dans le plus long bloc, quant à $\tau_0(n)$, elle est définie en fonction de n comme suit:

n	$n \leqslant 26$	$26 < n \leqslant 153$	$153 < n \leqslant 1170$
$\tau_0(n)$	$\tau_0 = 5$	$\tau_0 = 6$	$\tau_0 = 7$

Si l'une au moins des inégalités (11.65) est violée, l'hypothèse de stochasticité de l'échantillon doit être rejetée.

11.3.3. Test des carrés des différences séquentielles (test d'Abbe). Si l'échantillon $x_1, x_2, \ldots, x_n$ est extrait d'une population générale normale, on a intérêt à faire appel au test des carrés des différences séquentielles *) pour étudier sa stochasticité (l'hypothèse concurrente étant un déplacement systématique éventuel de la moyenne au cours du sondage).

Pour vérifier l'indépendance stochastique des observations à l'aide de ce test, on calcule la quantité

$$\gamma^{(n)} = \frac{q^2(n)}{s'^2(n)},$$

où

$$q^2(n) = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2;$$

$$s'^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$\bar{x} = \bar{x}(n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Si

$$\gamma(n) \leqslant \gamma_\alpha^{\min}(n), \tag{11.66}$$

l'hypothèse d'indépendance stochastique des observations est rejetée. Ceci étant, la quantité $\gamma_\alpha^{\min}(n)$ se calcule pour $n > 60$ à l'aide de la formule

$$\gamma_\alpha^{\min}(n) = 1 + \frac{u_\alpha}{\sqrt{n + 0{,}5(1 + u_\alpha^2)}},$$

où u_α est le quantile d'ordre α de la loi normale réduite.

Les quantités $\gamma_\alpha^{\min}(n)$ pour $n \leqslant 60$ sont données dans la table 4.9 [27] pour les trois valeurs du niveau α les plus souvent utilisées.

Exemple 11.6. Dans une cimenterie, on a prélevé quotidiennement durant 45 jours des échantillons pour déterminer la résistance moyenne à la compression des cubes de contrôle (N/cm^2 ou kg/cm^2).

---

*) Dans ce cas, ce test est plus puissant (cf. § 9.4) que le précédent. Cela signifie, en particulier, que si l'on se sert de ces deux tests pour un échantillon de taille n donnée et un risque de première espèce α donné (c'est-à-dire la probabilité de rejeter à tort l'hypothèse d'indépendance), le risque de deuxième espèce est plus petit pour le test des carrés des différences séquentielles.

Les résultats sont les suivants : 40, 33, 75, 18, 62, 33, 38, 69, 65, 100, 124, 91, 79, 42, 63, 23, 47, 52, 98, 97, 73, 85, 88, 40, 42, 51, 23, 75, 52, 126, 90, 111, 92, 109, 72, 28, 56, 17, 52, 68, 75, 102, 107, 77, 45.

Les calculs ont donné :

$$\sum_{i=1}^{45} (x_i - \bar{x})^2 = 37336\,; \quad s'^2 = 848{,}5\,;$$

$$\sum_{i=1}^{44} (x_{i+1} - x_i)^2 = 42819\,; \quad q^2(n) = 486{,}6\,; \quad \gamma(n) = 0{,}5735.$$

Prenons le niveau α égal à 0,05. Dans la table 4.9 [27], on trouve $\gamma_{0,05}^{\min}(45) = 0{,}7603$. Ceci dénote une « inadmissible petitesse » de la

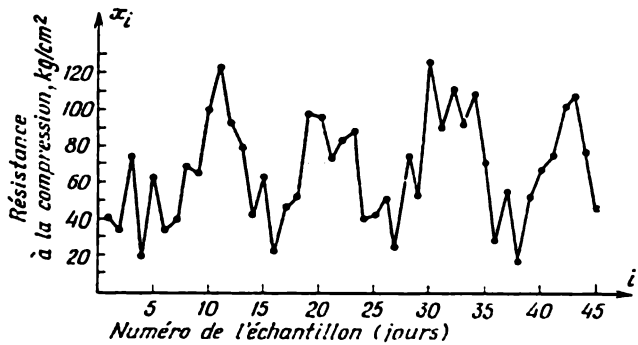


Fig. 11.1

quantité $\gamma(n)$, c'est-à-dire traduit la réalisation de l'inégalité (11.66). Il faut donc rejeter l'hypothèse d'indépendance stochastique des résultats des observations. La cause de cette « non-stochasticité » est probablement à chercher dans certaines tendances systématiques du comportement dans le temps de la moyenne de la variable aléatoire étudiée.

La figure 11.1 donne une idée suggestive du caractère cyclique de ces tendances.

11.4. Méthodes de traitement statistique des données manquantes

Dans ce paragraphe, on décrit les méthodes de traitement des matrices de données de la forme X dans le cas où une partie des mesures manque (cf. § 1.1). On admettra que l'absence d'une valeur d'un caractère de l'individu étudié (une colonne de la matrice des données) est liée à des causes techniques, par exemple, au mauvais

fonctionnement d'un instrument de mesure ou à une erreur grossière lors de la préparation des données, qui a occulté la véritable valeur du caractère, etc., mais pas à l'état de cet individu. Le chercheur peut avoir, selon la nature du problème, à estimer soit quelques paramètres dans le cas d'observations manquantes, soit ces observations manquantes, soit les deux à la fois. Les deux derniers cas exigent plus de conditions initiales que l'estimation des paramètres. Les méthodes de leur résolution utilisent une information redondante fournie par les relations entre les caractères.

Introduisons les définitions et notations suivantes. Appelons unité (colonne) *complète*, une unité dont on a mesuré toutes les variables. De façon analogue, appelons *variable complète* (colonne complète) une variable qui a été mesurée à l'intérieur de toutes les unités. L'ensemble des unités dont on a mesuré la variable x^i sera désigné par K_i, et le nombre de ces unités par n_i. On notera K_{ij} l'ensemble des unités dont on a mesuré les variables x^i et x^j, et n_{ij}, le nombre de ces unités.

Un procédé simple de traitement des données manquantes consiste à considérer la plus grande sous-matrice des données dont toutes les lignes et colonnes sont complètes. Lorsque l'échantillon est composé d'un nombre suffisant d'unités complètes, et que la reconstitution des données manquantes n'est pas l'objectif du traitement, ce procédé est le plus payant. Mais si les échantillons sont de petites ou moyennes tailles et que les mesures soient d'un coût élevé, il est naturel d'essayer d'utiliser toute l'information disponible.

11.4.1. Estimation des paramètres inconnus en cas d'observations manquantes. L'une des plus vieilles et des plus simples méthodes de traitement consiste à remplacer les observations manquantes de la variable x^i par la moyenne arithmétique calculée sur les autres données. La matrice complétée est ensuite utilisée comme à l'accoutumée, par exemple, pour estimer les éléments de la matrice des covariances (cf. (5.36')). L'estimation de la matrice des covariances ainsi obtenue sera visiblement à biais, en particulier, les éléments diagonaux (les variances) seront calculés par défaut. Le biais des variances peut être facilement éliminé par une estimation de ces variances uniquement sur les valeurs mesurées des variables correspondantes. D'autre part, le biais des éléments non diagonaux ne peut être pris en considération si des conditions supplémentaires ne sont pas imposées à la répartition des observations manquantes dans la matrice des données.

Citons un résultat allant dans ce sens, dû à V. Boulyguine [49].

Supposons que les valeurs manquantes de la variable x^i constituent un événement aléatoire statistiquement indépendant des valeurs prises par les autres variables de l'unité considérée et des valeurs de x^i à l'intérieur des autres unités. Soit $\tilde{s}_{ij}$ l'estimation d'un élément

de la matrice des covariances, obtenue par substitution des valeurs moyennes. L'estimation sans biais $\hat{s}_{ij}$ de l'élément σ_{ij} sera alors

$$\hat{s}_{ij} = \delta_{ij} \frac{\tilde{s}_{ij}(n-1)}{n_i - 1} + (1-\delta_{ij}) \frac{\tilde{s}_{ij}(n-1)^2}{(n_i-1)(n_j-1)}, \qquad (11.67)$$

où $\delta_{ij} = \begin{cases} 1 \text{ pour } i=j; \\ 0 \text{ sinon.} \end{cases}$

Mais les valeurs manquantes sont rarement indépendantes en pratique. C'est pourquoi il est plus sûr d'estimer le vecteur des moyennes et la matrice des covariances uniquement sur les données disponibles. Comme précédemment, pour estimer la moyenne et la variance de la variable x^i, on se sert de la moyenne arithmétique et de l'écart-type calculés sur les valeurs disponibles de cette variable, quant aux éléments non diagonaux de la matrice des covariances, ils sont estimés sur toutes les unités dont on a mesuré le *couple de variables*

$$\hat{m}_i = \frac{1}{n_i} \sum_{l \in K_i} x_l^i; \qquad (11.68)$$

$$s_{ij} = \frac{1}{n_{ij}-1} \sum_{l \in K_{ij}} (x_l^i - \hat{m}_i)(x_l^j - \hat{m}_j). \qquad (11.69)$$

Il est évident que pour estimer l'élément non diagonal s_{ij}, il faut disposer d'au moins deux unités dont on a mesuré le couple de variables x^i et x^j. L'estimation (11.69) est sans biais et converge si tous les $n_{ij} \to \infty$ avec n.

On observera toutefois que la précision de l'estimation (11.69) baisse rapidement lorsque le nombre des valeurs relevées croît. La matrice (11.69) n'est en général plus semi-définie positive. Bien plus, les coefficients de corrélation $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ peuvent être plus grands que 1 en module.

L'estimation de la matrice des covariances obtenue à l'aide de la procédure suivante est bien plus exacte même pour un grand nombre de données manquantes. Outre le coefficient de variation qui est donné par la formule (11.69), on calcule pour chaque couple de variables x_i et x_j $(i \neq j)$ les *estimations intermédiaires sur l'ensemble K_{ij} des couples de valeurs mesurées*

$$\hat{s}_l^2 = \frac{1}{n_{ij}-1} \sum_{r \in K_{ij}} (x_{rl} - m_l)^2 \quad (l = i, j),$$

puis l'estimation du coefficient de corrélation

$$\hat{r}_{ij} = s_{ij}/(\hat{s}_i \hat{s}_j).$$

Il est évident que $|\hat{r}_{ij}| \ll 1$ de par la procédure même de calcul de cette estimation. L'élément non diagonal $\hat{s}_{ij}$ s'estime maintenant de la manière suivante

$$\hat{s}_{ij} = \hat{r}_{ij} s_{ii} s_{jj}, \tag{11.69'}$$

où les éléments diagonaux s_{ii} sont estimés à l'aide de (11.69). Le calcul de l'estimation (11.69') nécessite une longue occupation de la machine et une mémoire supplémentaire pour les quantités $\hat{s}_i^2$ et $\hat{s}_j^2$, mais elle est bien meilleure que l'estimation (11.69) lorsque le nombre des données manquantes est élevé.

Les nombres de degré de liberté n_M^* et n_S^* correspondant aux estimations (11.68) et (11.69) constituent une caractéristique importante de leur certitude et de leur précision. Le nombre n_S^* peut être interprété comme la taille effective de l'échantillon sur le vu duquel est estimée la matrice des covariances, autrement dit, l'estimation (11.69) admet la même précision que l'estimation de la matrice des covariances, obtenue à l'aide d'un échantillon de taille n_S^* sans observations manquantes. La quantité n_M^* recouvre le même sens pour le vecteur des valeurs moyennes.

Dans [24] on propose d'utiliser les valeurs suivantes pour n_M^* et n_S^*:

$$n_S^* = \frac{p(p+1)}{2} \Big/ \sum_{i,j} \frac{1}{(n_{ij}-1)};$$

$$n_M^* = p \Big/ \sum_i 1/n_i, \tag{11.70}$$

c'est-à-dire les quantités inverses de la moyenne géométrique des unités de K_i et K_{ij}, où p est le nombre de variables. Les quantités (11.70) doivent être portées, par exemple, dans le test des hypothèses d'ajustement et d'homogénéité dans le cas vectoriel (cf. n° 11.2.7).

Contrairement à l'estimation standard (5.36'), les estimations (11.69) et (11.69') peuvent ne pas être semi-définies positives pour des échantillons de petite taille. Leurs valeurs propres peuvent, en particulier, être strictement négatives.

Les autres estimations de la matrice des covariances et du vecteur des moyennes, obtenues en reconstituant les données manquantes, sont étudiées au n° 11.4.3.

11.4.2. Utilisation des composantes principales. L'approche suivante est utilisée pour estimer aussi bien les observations manquantes dans la matrice des données que les valeurs des composantes principales d'une unité X_i non complète.

Supposons que les coefficients des vecteurs des composantes principales $U_1, U_2, \ldots, U_q$ aient été estimés par une méthode quel-

conque, par exemple, en tant que vecteurs propres des matrices (11.67) ou (11.69). Des propriétés des composantes principales, il s'ensuit alors que

$$X_i = \sum_{j=1}^{q} U_j z_i^j + \Delta_i, \tag{11.71}$$

où Δ_i est une variable aléatoire caractérisant l'erreur de la représentation (11.71); z_i^j les valeurs des composantes principales qui sont à estimer.

Dans de nombreux cas, les premières composantes principales ($q \leqslant p$) de (11.71) donnent lieu à une erreur Δ_i de norme petite.

Supposons maintenant que N_i est l'ensemble des numéros (indices supérieurs) des r_i variables mesurées de l'unité X_i. Pour estimer les composantes principales en n'utilisant que les variables mesurées de (11.71), on dispose du système de r_i équations linéaires à q inconnues:

$$x_i^l = \sum_{j=1}^{q} u_{lj} z_i^j \quad (l \in N_i).$$

Ce système se résout par la méthode des moindres carrés, ce qui nous amène au système d'équations normales

$$V^i Z_i = C_i,$$

où V^i est la matrice d'ordre q d'éléments $v_{st}^i = \sum_{l \in N_i} u_{ls} u_{lt}$, $s, t \in 1, \ldots$ $\ldots, q$; Z_i est l'estimation du vecteur des q premières composantes principales; C_i, le vecteur de dimension q de composantes

$$c_{ik} = \sum_{l \in N_i} u_{lk} x_i^l \quad (k = 1, \ldots, q).$$

Les vecteurs $U_1, \ldots, U_q$ étant deux à deux orthogonaux, la matrice V^i est manifestement non dégénérée si $r_i > q$, c'est-à-dire que le nombre de variables mesurées est plus grand que celui des composantes principales. Dans [64] on montre que si la matrice (11.69) intervient dans l'estimation des coefficients des vecteurs des composantes principales, les estimations des Z_i seront sans biais. En portant maintenant les estimations de Z_i dans (11.71), on peut obtenir les estimations des composantes manquantes du vecteur X_i.

11.4.3. Reconstitution des données manquantes et estimation des paramètres par la méthode du maximum de vraisemblance. Estimations du « point fixe ». Partageons la matrice X en deux parties Z et Y dont la répartition conjointe dépend d'un paramètre θ qui dans la suite désignera le vecteur des moyennes; Y est une matrice des covariances composée des unités complètes et des variables mesurées des unités incomplètes, Z l'ensemble des valeurs manquan-

tes. Supposons maintenant qu'on nous demande l'estimation $\hat{\theta}$ qui maximise le logarithme de la fonction de vraisemblance $l(\mathbf{Y}; \theta)$ pour $\mathbf{Y}$ fixe (cf. § 8.2). Cette estimation est difficile à obtenir par des calculs directs. En revanche, il est bien plus facile de trouver la valeur qui maximise le logarithme de la fonction de vraisemblance $l(\mathbf{Z}, \mathbf{Y}; \theta)$ si les valeurs manquantes sont reconstituées de façon arbitraire. Si maintenant on traite $\mathbf{Z}$ comme une variable aléatoire suivant une certaine loi de probabilité (dépendant de $\mathbf{Y}$), on peut trouver la valeur de θ qui maximise l'espérance mathématique de $l(\mathbf{Z}, \mathbf{Y}; \theta)$.

Considérons un procédé de réalisation de cette approche [19].

Supposons que $f(\mathbf{Z} \mid \mathbf{Y}; \theta)$ est la densité de la loi conditionnelle de $\mathbf{Z}$ sachant que $\mathbf{Y}$ et θ sont donnés et posons $l(\mathbf{Z} \mid \mathbf{Y}; \theta) \equiv \ln f(\mathbf{Z} \mid \mathbf{Y}; \theta)$. Alors

$$l(\mathbf{Z}, \mathbf{Y}; \theta) = l(\mathbf{Y}; \theta) + l(\mathbf{Z} \mid \mathbf{Y}; \theta). \tag{11.72}$$

Fixons maintenant une *valeur initiale* des paramètres θ_A, ce qui définira complètement la densité $f(\mathbf{Z} \mid \mathbf{Y}; \theta_A)$, et prenons l'espérance mathématique des deux membres de (11.72) en les intégrant avec la densité $f(\mathbf{Z} \mid \mathbf{Y}; \theta_A)$:

$$\int l(\mathbf{Z}, \mathbf{Y}; \theta) f(\mathbf{Z} \mid \mathbf{Y}; \theta_A)\, d\mathbf{Z} = l(\mathbf{Y}; \theta) + \int l(\mathbf{Z} \mid \mathbf{Y}; \theta) f(\mathbf{Z} \mid \mathbf{Y}; \theta_A)\, d\mathbf{Z},$$

ou, dans d'autres notations,

$$\mathbf{E}(l(\mathbf{Z} \mid \mathbf{Y}; \theta) \mid \mathbf{Y}; \theta_A) = l(\mathbf{Y}; \theta) + \mathbf{E}(l(\mathbf{Z} \mid \mathbf{Y}; \theta) \mid \mathbf{Y}; \theta_A).$$

Déterminons maintenant la valeur de θ_B qui maximise le premier membre de cette expression. La quantité θ_B dépend de θ_A, ce qu'on exprime par la relation

$$\theta_B = \psi(\theta_A).$$

Appelons *estimation du « point fixe »* de θ une valeur de θ, telle que

$$\theta = \psi(\theta), \tag{11.73}$$

et l'équation (11.73), équation du point fixe.

Signalons les principales propriétés des estimations du point fixe qui lient ces dernières aux estimations du maximum de vraisemblance à l'aide de $l(\mathbf{Y}; \theta)$.

1. Soit $\theta_C = \operatorname{Arg\,max}_{\theta} l(\mathbf{Y} \mid \theta)$, c'est-à-dire que θ_C est une estimation du maximum de vraisemblance, calculée sur les observations mesurées. Alors θ_C est solution de l'équation (11.73): $\theta_C = \psi(\theta_C)$, c'est-à-dire appartient à l'ensemble des estimations du point fixe.

2. Si $l(\mathbf{Z} \mid \mathbf{Y}; \theta)$ est une fonction dérivable, toute estimation du point fixe (c'est-à-dire toute solution de l'équation (11.73))

est soit un point de maximum, soit un point stationnaire de la fonction de vraisemblance $l(\mathbf{Y}; \theta)$.

Supposons maintenant que $\mathbf{X}$ est un échantillon extrait d'une population normale. Supposons, comme dans le n° 11.4.2, que N_i est l'ensemble des numéros des variables mesurées à l'intérieur des unités X_i, $\mathbf{Y}$, l'ensemble des valeurs connues des variables et

$$\theta = (M, \Sigma), \ \theta_A = (M_A, \Sigma_A), \ \theta_B = \psi(\theta_A) = (M_B, \Sigma_B).$$

Les équations du point fixe s'écrivent dans ce cas [19] :

$$\begin{cases} m_{jB} = \dfrac{1}{n} \sum\limits_{i=1}^{n} \hat{x}_{iA}^{j}; & (11.74) \\ \sigma_{jkB} = \dfrac{1}{n} \sum\limits_{i=1}^{n} (\hat{x}_{iA}^{j} - m_{jB})(\hat{x}_{iA}^{k} - m_{kB}) + \sigma_{jkA, N_i}]: \end{cases}$$

$$\begin{cases} \hat{x}_{iA}^{j} = \mathbf{E}(x_i^j \mid N_i; M_A, \Sigma_A); & (11.75) \\ \sigma_{ikA, N_i} = \operatorname{cov}\{x_i^j, x_i^k \mid N_i; M_A, \Sigma_A\}. & (11.76) \end{cases}$$

Pour déterminer les estimations par le maximum de vraisemblance (du point fixe), on choisit une valeur initiale de M_A et Σ_A (par exemple, les estimations (11.68) et (11.69)) et on organise une procédure cyclique jusqu'à ce que la différence entre les estimations obtenues itérativement ne soit plus significative.

Si x_i^j a été observée, alors $\hat{x}_{jA}^{i} = x_i^j$; si x_i^j est une valeur manquante, on peut l'estimer à l'aide de la quantité

$$\hat{x}_{iA}^{j} = \sum_{l \in N_i} b_l^{ij} x_i^l + a_i^j,$$

où b_l^{ij} est le coefficient de l'équation de régression linéaire de la variable x^j sur les valeurs mesurées du vecteur X_i ; a_i^j le terme constant de l'équation.

Le terme σ_{jkA, N_i} se calcule comme suit :

$$\sigma_{jkA, N_i} = \begin{cases} 0 & \text{si les variables } x^j \text{ et } x^k \text{ de } X_i \text{ sont mesurées ;} \\ -\sum\limits_{l \in N_i} b_l^{ij} \sigma_{ljA} & \text{si la variable } x^j \text{ de } X_i \text{ n'a pas été mesurée ;} \\ -\sum\limits_{l \in N_i} b_l^{ik} \sigma_{klA} & \text{si la variable } x^k \text{ de } X_i \text{ n'a pas été mesurée ;} \\ -\sum\limits_{l, r \in N_i} b_l^{ij} b_r^{ik} \sigma_{lrA} & \text{si les variables } x^j \text{ et } x^k \text{ de } X_i \text{ n'ont pas été mesurées.} \end{cases}$$

Pour conditions initiales, on peut prendre les valeurs de M et de Σ calculées à l'aide des expressions (11.68) et (11.69). Si, dans le schéma proposé, on se limite à une seule itération, on obtiendra le même résultat que par la méthode de reconstitution des données manquantes à l'aide d'une régression linéaire [25].

11.4.4. Méthode non paramétrique d'estimation des données manquantes. La méthode du point fixe, étudiée ci-dessus, implique que la loi de probabilité de la matrice des données soit exprimée sous une forme analytique, ce qui évidemment restreint sa portée. Cependant, il existe des méthodes de reconstitution des données manquantes qui n'exigent pas que soit connue la loi de probabilité, et qui utilisent la distance des unités (pour une certaine métrique), calculée à l'aide des variables mesurées à l'intérieur de ces unités. On admet que si deux unités sont proches dans l'espace des variables mesurées, elles le seront pour les variables non mesurées [142]. La métrique et la valeur de la distance qui détermine la proximité des unités dépendent des conditions du problème posé : des échelles dans lesquelles les variables sont mesurées, du nombre d'observations manquantes, etc. Le principe de l'une des éventuelles concrétisations de cette approche est le suivant. Supposons qu'il faut estimer la valeur manquante d'une variable $x^{(j)}$ de l'unité X_i, c'est-à-dire estimer l'élément $x_i^{(j)}$ de la matrice des données X. A cet effet, on considère la sous-matrice $X^{(j)}$ de X des colonnes (unités) dont sont mesurées les variables $x^{(j)}$, puis à l'intérieur de $X^{(j)}$ on envisage le groupe homogène des unités les plus proches de X_i dans l'espace des variables de X_i mesurées. On remplace ensuite la valeur manquante de $x_i^{(j)}$ par la moyenne de $x_i^{(j)}$ calculée sur le groupe homogène d'unités.

Cette procédure est réalisée par l'algorithme « ZET » qui est décrit en détail dans [142]. Les exemples traités *ibidem* montrent que cet algorithme fournit de bons résultats.

Considérons maintenant le problème de l'estimation de la qualité de la reconstitution des données manquantes par l'algorithme « ZET » et par les méthodes développées précédemment. A vrai dire, il est difficile d'introduire un critère formalisé de la qualité de la reconstitution des données manquantes. Mais il existe une méthode approchée d'estimation [142] qui consiste à exclure de façon aléatoire une partie des éléments de la matrice X et à reconstituer ensuite ces éléments à l'aide d'une méthode quelconque. L'écart (par exemple, la somme des carrés des écarts) entre les valeurs véritables et les valeurs reconstituées est précisément le critère de qualité de l'algorithme de reconstitution des observations manquantes.

11.5. Analyse des observations anomales

11.5.1. Position du problème. Dans ce numéro on étudie les méthodes d'identification des observations qui s'écartent fortement du centre de la répartition. Ces grands écarts sont le résultat d'une erreur aléatoire, d'une lecture incorrecte des indications de l'instrument de mesure, d'un décalage aléatoire de la virgule dans la représentation décimale de la mesure, etc., parfois le résultat de causes plus subtiles, telles qu'une non-conformité, en certains points, du modèle mathématique utilisé à la réalité, une modification des conditions de l'expérience qui a échappé au chercheur, etc. Quoi qu'il en soit, du point de vue mathématique, il s'agit de mettre en évidence les observations qui contrastent avec la principale masse des données. L'identification des observations anomales permet en général de revérifier les conditions dans lesquelles elles ont été enregistrées et traitées et, donc, de localiser et d'éliminer l'erreur. Si cette erreur ne peut être supprimée, c'est l'observation qui l'est, car étant non caractéristique.

Le problème posé se traite en deux étapes : identification des observations « suspectes » et test de la signification statistique de leurs écarts par rapport à la principale masse des données.

Il va de soi que les deux étapes sont basées sur certaines hypothèses portant sur la répartition aussi bien de la principale masse des données (la partie non polluée) que des « déchets ». On admet généralement que les observations de la partie non polluée suivent une loi normale à une ou plusieurs dimensions de paramètres inconnus. En analysant les écarts des observations par rapport au modèle mathématique, on postule parfois, de plus, que la moyenne de la répartition des écarts est nulle. Pour ce qui est des conditions imposées à la pollution, il n'y a pas d'archétype. On admet parfois que les déchets ont la même variance et une moyenne légèrement différente que le gros de l'échantillon, tout comme on admet qu'ils ont une moyenne peu différente de celle de la principale masse, mais une variance bien plus grande. Pour faciliter les références, nous allons formaliser ces conditions. Soient $x_1, x_2, \ldots, x_n$ un échantillon, $i_1, \ldots, i_{n_1}, j_1, \ldots, j_{n_2}$ $(n_1 + n_2 = n)$ des collections d'indices de l'ensemble $1, 2, \ldots, n$, correspondant aux parties non polluées et polluées de l'échantillon. La condition de non-pollution s'écrit :

$$x_i \in N(\mu, \sigma^2), \tag{11.77}$$

où μ et σ sont des paramètres inconnus.

Les conditions de pollution s'écrivent :

$$x_j \in N(\mu + d, \sigma^2) \text{ si la moyenne est déplacée ;} \tag{11.78}$$

$$x_j \in N(\mu, \gamma\sigma^2),\ \gamma \geqslant 1, \text{ si la variance est grande} \tag{11.79}$$

Si l'on sait *a priori* que la moyenne de la principale partie de l'échantillon est nulle, alors (11.77) devient

$$x_i \in N\ (0,\ \sigma^2), \tag{11.80}$$

et les relations (11.78), (11.79)

$$x_j \in N\ (d,\ \sigma^2); \tag{11.81}$$

$$x_j \in N\ (0,\ \gamma\sigma^2),\ \gamma \geqslant 1. \tag{11.82}$$

Avant de passer à la description des méthodes d'identification des déchets, signalons que l'approche purement statistique du problème d'identification et d'élimination des observations anomales, développée dans ce numéro, implique une certaine prudence au niveau de l'interprétation des données. L'hypothèse d'homogénéité, qui repose à la base des procédures statistiques, peut effectivement être mise en défaut et les déchets prennent un relief tout particulier en jetant la lumière sur le procédé de leur collecte.

11.5.2. Méthodes graphiques. Appelons *h-normal* (half-normal) un papier fonctionnel déduit du papier gausso-arithmétique de la manière suivante: on découpe le demi-plan inférieur correspondant aux ordonnées $<0{,}5$, puis on fait le changement d'ordonnées $t' = 2\ (t - 0{,}5)$. Sur le papier h-normal, la fonction de répartition de $|x|$, où $x \in N\ (0,\ \sigma^2)$ se représente par une droite issue de l'origine des coordonnées et de coefficient de direction $k = 1/\sigma$, c'est-à-dire le coefficient de direction de la fonction de répartition de x sur le papier gausso-arithmétique.

Considérons maintenant le cas où la condition (11.80) est imposée à la partie non polluée de l'échantillon et les conditions (11.81) ou (11.82), à la partie polluée. Si l'on construit la fonction de répartition de $|x_i|$ sur du papier h-normal, on obtient un graphique dont la partie gauche se laissera bien approcher par une droite issue de l'origine des coordonnées. La partie droite du graphique bifurquera vers la droite de cette droite et le point de bifurcation nous permettra d'estimer le taux de pollution.

Dans le cas général, si l'on a affaire aux modèles (11.80), (11.81), (11.82), on commence par estimer le paramètre μ (cf. § 10.4) par une méthode stable quelconque, puis on construit le graphique de la fonction de répartition de la différence $|x_i - \hat{\mu}|$ sur du papier h-normal. Le graphique admet les mêmes propriétés et la même interprétation que dans le cas particulier qui vient d'être traité.

Si, dans le cas multidimensionnel, on admet que la principale partie de l'échantillon suit approximativement une loi normale $N\ (X,\ M,\ \Sigma)$, on estime les paramètres de cette loi à l'aide des λ-moments (cf. § 10.4) pour une valeur de λ, telle que l'effet de la pollution sur l'estimation ne soit pas élevé et que les estimations obtenues possè-

dent de bonnes propriétés empiriques. On construit ensuite l'histogramme de la répartition

$$(X - M)\,\hat{\Sigma}^{-1}\,(X - \hat{M}). \tag{11.83}$$

Dans ces conditions, cette quantité doit suivre une loi du χ^2 à p degrés de liberté. L'allourdissement visible de l'extrémité droite de l'histogramme est le signe d'une pollution de l'échantillon. Ce procédé, même s'il est approximatif, est commode en pratique dans la mesure où le calcul des λ-moments pour chaque paramètre va de pair avec l'estimation, à l'aide de (11.83), du poids avec lequel l'observation figure dans les estimations $\hat{M}$ et $\hat{\Sigma}$.

11.5.3. Méthode analytique d'exclusion d'une observation extrême. Sans nuire à la généralité, on admettra qu'il est question de l'observation maximale. Soit $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$ la série variationnelle associée à l'échantillon. La règle de décision, relative à l'exclusion du terme extrême de la série variationnelle, est basée sur la statistique $T_n = (x_{(n)} - \bar{x})/s$, où $\bar{x}$ et s se définissent de la manière habituelle. La loi et les points de pourcentage de T_n ont été étudiés par E. Pearson [110], N. Smirnov [122], F. Grubbs [62]. La table des valeurs critiques de T_n établie par Grubbs est accessible dans [27]. Si T_n est plus grande que la valeur critique correspondante, l'hypothèse de la pollution est acceptée, sinon nous n'avons aucune raison du point de vue statistique de parler de pollution.

L'instabilité des estimations $\bar{x}$ et s (cf. § 10.4) aux écarts de la répartition de x_i par rapport à la loi normale réduit la portée pratique de ce test. En se servant des estimations stables des paramètres de localisation et d'échelle, on peut construire des tests de type T_n plus stables.

Si on suspecte l'existence de plusieurs valeurs extrêmes, on applique ce test à la plus grande d'entre elles. Si cette valeur est un déchet, on l'élimine de l'échantillon et on applique le test à la valeur suivante et ainsi de suite jusqu'à ce que l'on obtienne une valeur qui ne soit pas un déchet. Cette approche est compliquée par le fait que les observations suspectes forment un groupe compact à l'écart de la principale masse des observations, ce qui rend la procédure itérative basée sur l'utilisation de $\bar{x}$ et s insensible à ces observations. Comme au n° 11.5.2, on peut recommander de remplacer $\bar{x}$ et s par les λ-moments correspondants (cf. § 10.4 et [97], [98]).

11.5.4. Test analytique d'élimination simultanée de plusieurs observations extrêmes. Le test exposé plus bas a été proposé par G. Tietjen et H. Moore [132]. L'hypothèse nulle consiste comme toujours en ce que l'échantillon a été extrait d'une population normale. La règle d'élimination des k plus grands termes de la série variation-

nelle est basée sur la statistique

$$L_k = \frac{\sum_{i=1}^{n-k} (x_{(i)} - \bar{x}_k)^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2}, \tag{11.84}$$

où $\bar{x}_k$ est la moyenne des $n - k$ premiers termes de la série variationnelle, $\bar{x}$, la moyenne de l'échantillon tout entier. En cas de déchets, la statistique L_k doit être inférieure à la limite critique calculée pour une loi normale. Les tables des valeurs critiques pour L_k sont accessibles dans [124].

Si l'échantillon compte des déchets parmi les plus petites et les plus grandes observations, il faut modifier la règle d'estimation de leur signification, exposée ci-dessus. Cette modification est proche de celle décrite au n° 11.5.2 et consiste en ce qui suit. On calcule $\bar{x}$ sur les valeurs $x_1, \ldots, x_n$, puis les écarts absolus $| x_i - \bar{x} |$. On range ces écarts absolus par ordre de grandeur croissante et on les désigne par $z_{(i)}$. Soit $\bar{z}_{(k)}$ la moyenne arithmétique des $n - k$ premiers termes de la série variationnelle obtenue. Le test modifié s'écrit alors

$$\mathrm{E}_k = \frac{\sum_{i=1}^{n-k} (z_{(i)} - \bar{z}_k)^2}{\sum_{i=1}^{n} (z_{(i)} - \bar{z})^2}. \tag{11.85}$$

Le point faible de ce test est qu'il repose sur des statistiques qui dépendent fortement de la normalité de la loi et aussi que k n'est pratiquement jamais connu à l'avance mais estimé sur le vu des données, auxquelles sont appliquées les statistiques $\bar{x}$ et s. Cette dernière circonstance, ainsi que l'ont prouvé Tietjen et Moore, se répercute fortement sur le niveau critique réel du test, ce qui plaide une fois de plus pour les « naïves » méthodes graphiques.

Conclusions

1. Les tests d'hypothèses statistiques peuvent être répartis en deux classes : la classe des tests dont la répartition des statistiques critiques (sachant que l'hypothèse nulle est vraie) dépend de celles de l'échantillon (ces tests sont dits paramétriques), la classe des tests dont les statistiques sont « distribution free », c'est-à-dire les tests dont la répartition des statistiques critiques (sachant que l'hypothèse nulle est vraie) ne dépend pas de la répartition de la population générale (ces tests sont dits non paramétriques). Cette classification des tests est jusqu'à un certain point conventionnelle. Ainsi, dans le cas

d'un test du χ^2, la répartition de la statistique critique, sachant que l'hypothèse nulle est vraie, ne dépend pas de la répartition modèle, bien que pour l'appliquer on ait généralement besoin de l'estimation des paramètres de cette répartition modèle. D'autre part, l'utilisation des tests non paramétriques de Kolmogorov et du ω^2 dépend déjà de la forme de la répartition modèle si elle implique l'estimation des paramètres de cette dernière.

2. Pour tester l'hypothèse de normalité de la répartition modèle, on peut faire appel aux tests traditionnels basés sur les valeurs empiriques des coefficients de dissymétrie et d'aplatissement. L'application de ces tests est compliquée par la convergence lente des répartitions des statistiques critiques vers les répartitions limites, d'où la nécessité d'utiliser des tables de points de pourcentage des répartitions exactes des statistiques critiques, dressées pour des échantillons de tailles fixes. Les statistiques des tests de Kolmogorov et du ω^2, destinés à vérifier la normalité d'une répartition dans le cas où les paramètres de cette dernière sont estimés sur le vu de l'échantillon, tendent bien plus vite vers les répartitions limites.

3. Le test du χ^2 est utilisé pour éprouver les hypothèses d'ajustement et d'homogénéité si les données sont groupées. La répartition limite de la statistique de ce test, sachant que l'hypothèse nulle est vraie, ne dépend pas de la répartition de la population générale, bien que ce test implique l'estimation des paramètres de la répartition modèle. Un point capital dans ce test est le choix du nombre d'intervalles de groupement et la répartition des observations sur ces intervalles.

4. Si la répartition modèle est entièrement connue et est continue (cette situation se présente lorsqu'on vérifie, par exemple, des générateurs de nombres aléatoires suivant une loi de probabilité donnée), les tests de Kolmogorov-Smirnov et Cramer-Mises sont les plus appropriés pour tester l'hypothèse d'ajustement. Les répartitions des statistiques de ces tests convergent vers la répartition limite, quant aux répartitions limites, elles se calculent aisément. Les répartitions des modifications des statistiques critiques convergent encore plus vite vers les répartitions limites. Dans le cas de données groupées, le niveau de ces tests sera inférieur au niveau nominal.

5. Pour vérifier l'hypothèse d'homogénéité de deux échantillons en dimension un, on peut se servir des tests de Smirnov et des tests de rang linéaires. Le test de Smirnov est efficace pour toute violation de l'hypothèse nulle, mais n'indique pas la nature de cette violation. Les tests de rang linéaires permettent de dire à quoi est due l'hétérogénéité : à l'inégalité des paramètres de localisation et ou d'échelle. Le mérite des tests de rang linéaires réside dans la convergence rapide des répartitions de leurs statistiques vers la répartition limite (normale) et leur stabilité à la pollution. Les tests de rang linéaires se

généralisent facilement au cas de plus de deux classes. Ces tests ne s'appliquent tous à strictement parler qu'à des répartitions modèles continues. L'application des tests de rang linéaires aux données groupées implique l'introduction de corrections appelée traitement des coïncidences.

6. Si l'on sait que les échantillons sont prélevés dans des populations normales, on peut, pour en tester l'homogénéité, se servir de tests basés sur les statistiques bien connues de Student et de Fisher. Ces tests se généralisent à un nombre k (>2) de classes. Le test basé sur la statistique de Student est stable pour les écarts par rapport à la loi normale et peut être utilisé dans une situation plus générale.

7. Pour vérifier l'homogénéité dans le cas multidimensionnel, on ne dispose en fait que de tests qui partent de l'hypothèse que les échantillons sont extraits de populations normales multidimensionnelles. Le test du T^2 est comme son analogue unidimensionnel, le test t de Student, stable pour les écarts par rapport à la loi normale et peut être utilisé dans une situation générale.

8. Pour éprouver l'hypothèse de symétrie d'une répartition, on peut solliciter des tests de rang linéaires qui sont analogues aux tests de rang linéaires de l'hypothèse d'homogénéité.

9. Les procédures statistiques d'identification des observations anomales conjecturent l'homogénéité des données. Ceci étant, les déchets sont traités comme des observations qui sont anormalement éloignées du centre de la répartition. De nombreuses procédures analytiques d'identification des déchets et d'estimation de la signification de leurs écarts ont été proposées ces derniers temps. La principale difficulté de leur mise en œuvre réside dans le fait que le taux de « pollution » n'est pas connu et s'estime sur ces données mêmes qui sont utilisées pour tester la signification des écarts. Les procédures graphiques sont les plus stables aux écarts par rapport à l'hypothèse de normalité de la principale partie de l'échantillon.

Signalons que les déchets peuvent fournir de précieux renseignements sur la manière dont les données ont été recueillies.

CHAPITRE 12

LOGICIEL DE LA STATISTIQUE APPLIQUÉE ET QUELQUES PROBLÈMES RELATIFS À LA TECHNIQUE DE CALCUL

12.1. Logiciel de la statistique appliquée

Le logiciel de la statistique appliquée (SA) est bien élaboré et se développe intensément aussi bien dans le sens de la conception de nouvelles méthodes de traitement des données, que du perfectionnement de son organisation et de sa structure. Les *paquets de programmes* *) sont une forme commode d'organisation du logiciel, qui permet de traiter efficacement les données. Citons les critères essentiels de qualité d'un paquet de programmes de SA.

Le premier groupe de critères a trait à l'accès et à la commodité de la gestion du paquet pour l'utilisateur. S'y rapportent :

1. Une *documentation sur le paquet* qui soit détaillée et bien présentée. Il s'agit en premier chef d'un « Manuel d'utilisation » destiné à initier l'utilisateur aux divers usages du paquet et à la façon de s'en servir. Ce manuel est indispensable à la diffusion du paquet de programmes. Comme exemples de manuels bien présentés citons ceux des paquets *BMDP* [24], [25] et *SPSS* [125].

2. La possibilité de *formuler le problème* et de *gérer les données* à l'aide d'un langage simple adapté à la statistique appliquée.

3. La mise en place d'un système *d'indication et de commentaire détaillé des erreurs* commises lors de la formulation du problème et dans le processus de traitement.

4. L'instauration d'un *régime conversationnel avec le paquet.*

L'autre groupe de critères est lié aux qualités fonctionnelles du paquet.

5. Des méthodes accessibles de manipulation des données (triage, rédaction, transformation, agrégation et séparation des données, etc.) avant leur traitement par les méthodes de SA.

6. Un arsenal assez riche de procédures de SA pour le traitement des données.

7. La possibilité d'organiser des procédures itératives séquentielles de traitement des données.

8. La possibilité de traiter des données de grande dimension.

Le troisième groupe de critères est important pour le traitement prolongé de grands fichiers.

---

*) Le développement du logiciel de SA va dans le même sens que celui du logiciel appliqué en général.

9. L'existence de méthodes de contrôle et de stockage des données.
10. La possibilité de relier le paquet aux banques de données.

12.1.1. Organisation des paquets de programmes. On distingue les niveaux suivants d'organisation des paquets de programmes.

Au premier niveau on trouve le *paquet de structure simple* qui est une bibliothèque de modules : par modules on entend une procédure, un sous-programme ou un programme rédigé dans un langage de programmation évolué (FORTRAN, PL/1), satisfaisant à des contraintes assurant l'unité conceptuelle du paquet. Les plus importantes d'entre elles sont les contraintes qui portent sur la structure de la mémorisation des principales unités d'information, nécessaires à la compatibilité des entrées et des sorties des divers modules. Il faut, par exemple, que la matrice des données soit présentée à l'entrée des modules sous forme d'un tableau à deux dimensions dont les éléments sont disposés en lignes (ou en colonnes). Les autres contraintes sont liées aux moyens de transmission de l'information (par exemple, l'admission ou l'interdiction de variables exogènes), à l'utilisation d'opérateurs de type entrée-sortie, à l'achèvement en catastrophe du travail du module. Les sous-programmes de ces paquets sont dotés aussi d'une documentation compatible. Le paquet de structure simple peut être utilisé aussi bien sous forme d'une bibliothèque des modules initiaux (dans un langage de programmation évolué) que sous forme d'une bibliothèque de modules de chargement obtenus par compilation et rédaction des modules initiaux respectifs. Un exemple de paquet de structure simple est le Paquet des sous-programmes scientifiques (PSS) écrit en FORTRAN pour le Système Unique des ordinateurs [88] présenté généralement sous forme de modules de chargement. Les sections des PSS qui nous intéressent sont décrites dans [vyp. 2], [vyp. 10]. Un PSS a été composé en PL/1 (cf. [vyp. 14]). Le paquet *SSP* de la firme IBM a servi de modèle pour la version du PSS en FORTRAN, la version *SSP* sur PL/1, respectivement pour la version PSS en PL/1. L'arsenal des procédures de statistique appliquée dans le PSS s'est considérablement enrichi en comparaison des paquets initiaux *SSP*, notamment dans le domaine du traitement des données multidimensionnelles. En plus des programmes purement statistiques, ces paquets contiennent des sous-programmes empruntés à d'autres branches du calcul numérique : algèbre linéaire, interpolation et extrapolation des fonctions tabulées, etc.

L'utilisation du paquet de structure simple implique de l'utilisateur une bonne maîtrise de la programmation. L'utilisateur doit sélectionner les modules appropriés et rédiger le programme-maître en un langage évolué qui assure l'appel des modules sélectionnés dans un ordre déterminé, le chargement et la gestion des entrées-sorties. A ce niveau le langage de l'utilisateur comprend un langage de programmation évolué (FORTRAN, PL/1), un langage d'appel du

système d'opérations et des procédures linguistiques introduites par l'utilisateur pour gérer son programme.

A un niveau bien plus élevé d'organisation, on trouve les paquets constitués d'un ensemble de programmes prêts à l'usage, commandés par un langage d'entrée et possédant la même entrée-sortie des données. A ce niveau d'organisation il est possible d'élaborer des méthodes relativement simples de manipulation, de contrôle et de stockage des données.

Les programmes du paquet sont constitués de modules qui en principe sont pris dans une bibliothèque organisée comme nous l'avons décrit plus haut. Les programmes comportent généralement des overlays *) pour réduire les ressources de la mémoire principale.

L'utilisateur n'a pratiquement aucune possibilité de changer les programmes, hormis les cas, prévus à l'avance, de modification du volume de la mémoire réservée aux données et la connexion, suivant des règles bien définies, de certaines procédures de l'utilisateur en FORTRAN ou PL/1, par exemple des procédures de conversion des variables. Cependant ces paquets permettent de satisfaire l'écrasante majorité des besoins de l'utilisateur grâce à la souplesse du langage d'entrée et la possibilité d'élaborer sans peine des procédures en cascade de traitement avec les programmes du paquet. Dans le même temps, ces paquets agissant eux-mêmes sur le déroulement de l'analyse statistique en conditionnant, par la composition des procédures permises et les contraintes imposées aux données d'entrée, la forme de la préparation, et dans une grande mesure, de la collecte des données. La manipulation d'un tel paquet est à la portée d'un utilisateur profane en programmation, bien qu'en règle générale un minimum de connaissances des langages de communication avec le système d'opérations soit requis.

L'un des plus connus des paquets de ce type est le *BMDP* mis au point aux U.S.A. Nous envisagerons plus bas deux versions de ce paquet : la version de 1975 [24] adaptée en U.R.S.S. pour le SU des ordinateurs et la version de 1979 [25] qui a été considérablement plus étendue par rapport à celle de 1975. Dans la suite, nous désignerons ces versions par *BMDP* 75 et *BMDP* 79.

Un autre paquet de ce type est le paquet de Programmes d'analyse statistique appliquée (PPSA) [108], élaboré par l'Institut central d'économie et de mathématiques de l'Académie des sciences de l'U.R.S.S. Le PPSA diffère du *BMDP* par la composition des méthodes de traitement réalisées, le langage d'entrée, l'organisation et la documentation des données, ainsi que par l'organisation des programmes. Si dans le *BMDP* chaque programme ne réalise qu'une seule méthode de traitement, le PPSA utilise les programmes thé-

*) L'organisation d'une structure overlay permet de charger les modules de traitement dans la mémoire au moment où ils commencent à traiter les données et de les remplacer par des modules qui ont fini de fonctionner (cf. [86]).

matiquement orientés (programmes TO) [10]. Chaque programme TO est en fait un paquet de programmes pour la résolution d'un sous-ensemble de problèmes de statistique appliquée (estimation, analyse de régression, discriminante, etc.). Les sous-ensembles de modules sollicités par les divers programmes TO se recoupent partiellement. C'est ainsi que les programmes exploitent tous les mêmes modules d'entrée et de traduction des propositions du langage d'entrée, d'introduction des données, de délivrance des données sous forme de tables et graphiques, etc. L'organisation des liaisons entre les modules dans les programmes TO permet de mettre au point une structure overlay simple. L'usage des programmes TO facilite dans bien des cas la tâche de l'utilisateur au niveau de la conception des procédures en cascade de traitement.

L'étape suivante de l'évolution des programmes TO (du point de vue organisationnel mais pas chronologique) est l'élaboration de paquets de programmes dont les procédures de traitement sont basées sur l'analyse des propositions du langage d'entrée décrivant le problème de l'utilisateur. Les méthodes de réalisation de cette approche sont très variées. Les paquets étrangers *P-STAT* [113], *SPSS* [125] et les paquets soviétiques SOD-GS [109], OTEKS [106] font partie de ce type.

Apparemment, pour l'utilisateur, le travail avec des paquets de ce type se distingue peu du travail avec les paquets *BMDP* ou PPSA. Aussi, dans la suite, nous ne ferons aucune distinction entre eux et nous les désignerons simplement par *paquets de programmes* en les distinguant des paquets (bibliothèques) de sous-programmes de type PSS. Tous les paquets mentionnés ci-dessus sont destinés au SU des ordinateurs ou à IBM-380/370 *) exception faite de la version PSS en FORTRAN qui peut être utilisée (sous forme de modules initiaux) sur les ordinateurs BESM-6 et « Minsk-32 ». Parmi les paquets de programmes destinés à l'ordinateur BESM-6, signalons le paquet SORRA-1 **) rédigé pour la résolution des problèmes de classification multidimensionnelle et de régression, et le paquet *DIAS* [53].

12.1.2. Problèmes d'organisation et possibilités de gestion des données. Les données d'entrée traitées par les paquets de programmes de statistique appliquée sont présentées essentiellement sous la forme d'une matrice de type « objet-variable ». Pour tous les paquets de programmes envisagés, on admet que la matrice des données est ordonnée comme suit :

$$x_1^{(1)}, \ldots, x_1^{(p)}, x_2^{(1)}, \ldots, x_2^{(p)}, \ldots, x_n^{(1)}, \ldots, x_n^{(p)}. \qquad (12.1)$$

*) Certains des paquets mentionnés peuvent être transposés sur un miniordinateur de type SM-4.

**) Cf. : Raoudis Ch. *Algorithmes de classification et de régression du système SORRA-1*. — Dans l'ouvrage : « Problèmes statistiques de gestion ». Vilnius, 1978, vyp. 77 (en russe).

Les données peuvent être introduites soit par le biais de cartea perforées (ou de leurs images sur bande magnétique et disque magnétique), soit par le biais de fichiers standards (pour le paquet envisagé) sur disque magnétique ou bande magnétique, ces fichiers représentant l'information sur la matrice des données en binaire ligne par ligne ou colonne par colonne, de telle sorte que chaque enregistrement est une unité *). Les variables de la matrice des données peuvent être quantitatives, qualitatives (ordinales) ou nominales (cf. chap. 10). Certaines valeurs peuvent ne pas être déterminées du tout. Mais dans les programmes des paquets, tous les éléments de la matrice des données sont supposés être des nombres. Ces nombres seront donc des codes conventionnels pour les variables non quantitatives. Si l'information initiale est donnée sous une forme symbolique, le paquet PPSA prévoit un programme spécial de recodage de cette information sous une forme numérique définie par l'utilisateur.

Considérons maintenant les possibilités de gestion des données offertes à l'utilisateur par les ressources insérées dans les paquets de programmes pour l'exécution des fonctions suivantes:

1. Mémorisation des sorties sous forme de fichiers standards, par exemple, de matrices des données transformées, de matrices des covariances, de matrices des proximités, etc.

2. Documentation des entrées et des sorties par des codes de fichiers ou des fichiers d'archives, etc.

3. Rédaction, c'est-à-dire complètement et correction des fichiers standards.

Tableau 12.1

Fonction / Paquet	Introduction des données sur cartes perforées	Introduction des données par fichiers standards sur bande et disque magnétiques	Mémorisation des données de sortie sous forme de fichiers standards	Rédaction des fichiers standards	Utilisation de la matrice des corrélations pour données d'entrée	Identification des données de sortie
BMDP 75	oui	oui	oui	non	oui	oui
BMDP 79	oui	oui	oui	non	oui	oui
SPSS	oui	oui	oui	non	oui	oui
PPSA	oui	oui	oui	oui	oui	oui
SOD-GS	oui	non	non	non	oui	non
OTEKS	oui	oui	oui	oui	non	oui
SORRA-1	non	oui	non	non	non	non
DIAS	oui	oui	oui	non	non	oui

*) L'organisation des données sur bande et disque magnétiques pour les paquets de programmes destinés à l'ordinateur BESM-6 est conditionnée par le système d'opérations correspondant.

Toutes ces informations sont récapitulées avec d'autres données sur le tableau 12.1 pour les paquets de programmes énumérés ci-dessus.

12.1.3. Méthodes de traitement préliminaire (manipulation) des données. Avant d'entamer le traitement statistique proprement dit des données, il faut réaliser les opérations suivantes:

1. Déterminer le sous-ensemble des variables, défini par une liste de noms ou de numéros.
2. Déterminer le sous-ensemble des objets, défini par une liste de noms ou de numéros.
3. Déterminer les sous-ensembles des lignes ou colonnes des autres matrices (des covariances, des proximités), définis par des listes de noms ou de numéros s'ils sont utilisés comme unités d'entrée.
4. Déterminer le sous-ensemble des objets dont les valeurs des variables se trouvent à l'intérieur (ou à l'extérieur) d'un domaine donné.
5. Grouper les objets:

a) selon que la valeur de la variable donnée, appelée généralement variable de groupement, appartient à tel ou tel intervalle;

b) selon d'autres critères de groupement.

6. Transformer les variables:

a) par des transformations définies dans le langage d'entrée;

b) par la procédure par l'utilisateur.

Tableau 12.2

Numéro du traitement / Paquet	1	2	3	4	5		6		7
					a	b	a	b	
BMDP 75, 79	oui	oui	oui	oui	oui	oui	oui	oui	oui
SPSS	oui	oui	oui	oui	oui	oui	oui	oui	oui
PPSA	oui	oui	oui	oui	oui	oui	oui	oui	oui
SOD-GS	oui	non	non	non	non	non	non	non	oui
OTEKS	oui	oui	non	oui	non	oui	non	non	non
PSS	oui	oui	oui	oui	non	oui	oui	—	—
SORRA-1	oui	non	non	non	non	oui	oui	—	non
DIAS	oui	non	non	oui	non	oui	non	oui	non

7. Normaliser les éléments de la matrice des données (à l'aide des écarts-types, de l'étendue, etc.).

Ces opérations sont représentées pour les divers paquets sur le tableau 12.2.

12.1.4. Traitement des données manquantes. Voyons les possibilités de traitement d'un fichier incomplet. Les méthodes auxquel

les nous ferons référence ont été décrites essentiellement au § 11.4. Considérons les méthodes suivantes de traitement des données manquantes :

1. Codage des données manquantes par des codes spéciaux de données manquantes.
2. Suppression des objets incomplets.
3. Estimation de la matrice des covariances et du vecteur des moyennes :

a) à l'aide des formules (11.69), c'est-à-dire en tenant compte de toutes les valeurs relevées des couples de variables (pour les éléments non diagonaux de la matrice des covariances) et de toutes les valeurs relevées de la variable pour l'estimation de la moyenne et de la variance ;

b) à l'aide d'autres méthodes n'impliquant pas la reconstitution des données manquantes.

4. Reconstitution des données manquantes :

a) par les composantes principales ;

b) par une régression linéaire sur les variables mesurées ;

c) par d'autres algorithmes (par exemple, l'algorithme ZET) ;

d) par les moyennes.

5. Autres possibilités de traitement des fichiers incomplets.

Les paquets SOD-GS, SORRA-1 ne contiennent pas de procédures de reconstitution des données manquantes. Par contre, pour le paquet OTEKS cette opération fait partie de ses plus importantes fonctions et elle est réalisée à l'aide de l'algorithme ZET. Les méthodes de traitement des fichiers incomplets sont regroupées dans le tableau 12.3.

Arrêtons-nous en détail sur certaines méthodes auxiliaires de reconstitution des données manquantes réalisées dans les paquets

Tableau 12.3

Numéro de traitement des données manquantes (1)-(5) / Paquet	1	2	3		4				5
			a	b	a	b	c	d	
BMDP 75	oui	oui	oui	oui	non	oui	non	oui	non
BMDP 79	oui	oui	oui	oui	non	oui	non	oui	oui
SPSS	oui	oui	oui	non	non	oui	non	oui	non
PPSA	oui	oui	oui	non	oui	non	non	oui	non
OTEKS	oui	oui	non	non	non	non	oui	non	oui
PSS	oui	oui	oui	non	non	non	non	non	non
DIAS	oui	oui	non	non	non	non	non	non	non

BMDP 79 et PPSA. Comme indiqué au § 11.4, l'estimation de la matrice des corrélations (des covariances) obtenue par la méthode 1 (cf. formule (11.69)) peut ne pas être semi-définie positive. Les paquets *BMDP* 79 et PPSA prévoient une procédure de vérification de la semi-définition positive de la matrice **S** par le calcul de ses valeurs propres (ou des valeurs propres de la matrice des corrélations correspondante). Si certaines valeurs propres sont strictement négatives, on peut obtenir une estimation semi-définie positive de la matrice des corrélations (des covariances) par un « lissage » qui consiste d'abord à calculer toutes les valeurs et vecteurs propres de la matrice des corrélations **B** obtenue et à construire une matrice $\mathbf{A} = \mathbf{U}'\mathbf{L}\mathbf{U}$, où **U** est la matrice des vecteurs propres associés aux valeurs propres *strictement positives* de la matrice **R**, et **L** la matrice diagonale des *valeurs propres strictement positives*. En normant ensuite la matrice **A**, on obtient la matrice des corrélations $\mathbf{R}^*$. Si, par la suite, on a besoin de la matrice des covariances **S**, on peut la déduire de $\mathbf{R}^*$ en multipliant les colonnes et les lignes par les estimations des écarts-types. Il est évident que $\mathbf{R}^*$ comme $\mathbf{S}^*$ seront des matrices semi-définies positives de rang incomplet. Toutefois, étant donné que le rang de ces matrices et les valeurs et vecteurs propres de **R** sont connus, il est aisé de déterminer la matrice inverse généralisée (cf. [119]) pour les besoins de l'analyse de régression et discriminante.

12.1.5. Traitement primaire des données qualitatives. Considérons les méthodes suivantes de traitement des données qualitatives:

1. L'analyse des tableaux de contingence comprend

1.1) une estimation des paramètres et un test des hypothèses d'indépendance pour les divers modèles de tableaux de contingence;

1.2) une estimation des paramètres d'un modèle loglinéaire du tableau de contingence;

1.3) le calcul des divers coefficients de liaison dans les tableaux de contingence à deux dimensions.

2. Codage des variables en binaire (n° 10.2.4).

3. Codage des variables qualitatives.

Les tableaux de contingence (1.1 et 1.3) sont analysés dans les paquets *BMDP* 75 et *BMDP* 79 (les programmes *P9D*, *P1F*), et dans le paquet PPSA, d'après 1.3. Le paquet *BMDP* 79 (programme *P7D*) est muni d'une procédure d'estimation des paramètres du modèle loglinéaire.

Les problèmes de classification sont résolus en binaire dans le paquet OTEKS. Dans les paquets *BMDP* 75, *BMDP* 79 et PPSA, on passe facilement aux variables binaires, bien qu'il n'existe pas de programme spécial pour cette conversion.

Décrivons plus en détail la méthode de *codage des variables qualitatives*. Cette méthode consiste à affecter aux modalités des variables qualitatives des codes « raisonnables » dans le contexte du problè-

me envisagé. Les variables codées sont ensuite manipulées comme des variables numériques discrètes. Le critère d'affectation des codes dépend du traitement statistique ultérieur: analyse en composantes principales, analyse de régression, analyse discriminante. Le codage des variables ordinales semble assez naturel. Le codage des variables nominales est basé sur le fait qu'aux modalités desdites variables correspond un ensemble de valeurs de variables sous-jacentes, qui ne peuvent être mesurées directement mais qui existent réellement. De sorte que le code attribué à une modalité d'une variable nominale est une valeur résultante d'un ensemble de valeurs de variables non mesurées, caractérisant les objets correspondant à la modalité envisagé de cette variable nominale.

Les critères régissant le codage (cf. [49]) dépendent de la méthode d'analyse statistique qui sera mise en œuvre. Cependant, ce sont tous des fonctionnelles de la matrice des covariances (des corrélations) dans l'espace des variables codées. Ceci est lié tout d'abord au fait que la matrice des covariances (des corrélations) est le principal instrument utilisé par les méthodes d'analyse statistique énumérées plus haut.

Ces méthodes consistent à mettre en évidence une ou plusieurs combinaisons linéaires des variables initiales jouissant de certaines propriétés extrémales; quant à leur succès, il dépend essentiellement de l'intensité des liens entre les variables et du degré avec lequel la matrice des covariances reflètera ces liens, c'est-à-dire de la précision avec laquelle on peut considérer que ces relations sont linéaires.

Par exemple, si après le codage on a l'intention d'utiliser les méthodes de réduction de la dimension ou d'analyse de la régression dans le programme CODAGE du paquet PPSA, on affecte aux modalités des variables qualitatives des codes numériques maximisant le critère

$$R^2 = \sum_{i<j} \rho_{ij}^2 w_{ij}, \tag{12.2}$$

où $i, j = 1, \ldots, p$, p étant le nombre de variables à coder; ρ_{ij} le coefficient de corrélation entre la i-ième et la j-ième variable après codage, w_{ij} des coefficients de pondération positifs. La procédure de calcul est la suivante.

Soit Z_i, $i = 1, \ldots, p$, un vecteur de dimension L_i (L_i est le nombre de modalités de la i-ième variable) dont les coefficients sont les codes affectés aux modalités correspondantes de la i-ième variable.

Estimons au vu de l'échantillon de base la matrice (le tableau) de contingence $\mathbf{Q}(i, j)$ de format $L_i \times L_j$ ($i = 1, \ldots, p$; $j = 1, \ldots, p$) dont la valeur de l'élément q_{kl} est une estimation de la probabilité que la i-ième et la j-ième variable prennent simultanément, la première, la modalité k, la deuxième, la modalité l. Soient P une matrice diagonale dont les éléments sont les fréquences des

modalités de la i-ième variable, $\mathbf{A}_i$ une matrice symétrique semi-définie positive d'éléments

$$a_{lk}^{(i)} = \sum_{i<j} (\mathbf{Q}'_{l\bullet}(i, j), Z_j)(\mathbf{Q}'_{k\bullet}(i, j), Z_j), \tag{12.3}$$

où l et k varient de 1 jusqu'au nombre de modalités de la i-ième variable ; $(\mathbf{Q}'_{l\cdot}(i, j), Z_j)$ le produit scalaire de la l-ième ligne de la matrice $\mathbf{Q}(i, j)$ par le vecteur Z_j dont les composantes sont les codes des modalités correspondantes de la j-ième variable.

Les codes numériques qui maximisent le critère (12.2) sont déterminés par la procédure itérative suivante.

Au premier pas, on se donne les valeurs initiales de $Z_1^{(0)}, \ldots, Z_p^{(0)}$ (par exemple, les coordonnées de chaque vecteur $Z_i^{(0)}$, $i = 1, \ldots, p$, sont des entiers naturels, c'est-à-dire les numéros des modalités de la i-ième variable). Ces codes sont normés et centrés. On calcule ensuite la matrice $\mathbf{A}_1$ à l'aide de la formule (12.3) et on trouve le vecteur propre associé à la plus grande valeur propre pour le problème

$$\mathbf{A}_1 Z_1 - \lambda_1 \mathbf{P}_1 Z_1 = 0. \tag{12.4}$$

Les coordonnées de ce vecteur seront les nouvelles valeurs des codes pour $Z_1^{(1)}$.

Sachant $Z_1^{(1)}$, on calcule la matrice $\mathbf{A}_2$ pour $Z_1^{(1)}, Z_3^{(0)}, \ldots, Z_p^{(0)}$ fixes et on trouve un nouveau vecteur $Z_2^{(1)}$. On calcule de même $Z_3^{(1)}$, $Z_4^{(1)}, \ldots$

Une fois qu'on connaît toutes les valeurs des codes $Z_1^{(1)}, \ldots, Z_p^{(1)}$, on passe à la détermination de $Z_1^{(2)}$ pour $Z_2^{(1)}, \ldots, Z_p^{(1)}$ fixes et on relance les calculs.

Cette procédure prend fin lorsque la différence entre deux itérations successives est inférieure à un seuil donné.

Remarque. Lorsque le nombre de variables $p = 2$, le programme nous conduit aux codes de Lankaster [77].

12.1.6. Méthodes de visualisation des données. Considérons le logiciel des méthodes suivantes de visualisation des données (cf. § 10.5) :

1) méthode standard des composantes principales ;
2) méthode non linéaire des composantes principales ;
3) analyse des proximités multidimensionnelle linéaire.

La méthode standard des composantes principales est réalisée dans les paquets *BMDP* 75, *BMDP* 79 (dans le programme d'analyse factorielle *P4F*), dans le PSS [98, vyp. 2, 14], dans le PPSA (programme *REDUCT*). La méthode non linéaire des composantes principales et l'analyse des proximités linéaire sont réalisées seulement dans le PPSA (le programme *REDUCT*).

L'efficacité des méthodes de visualisation se juge à la possibilité de délivrer le diagramme de dispersion (scattergramme) (cf. chap. 10)

par toute sorte de terminaux : imprimante, traceur de courbes ou écran. L'imprimante est un instrument assez commode qui est passé dans l'usage.

Voyons plus en détail comment est construit le diagramme de dispersion (par exemple, dans le PPSA). Les coordonnées des projections des points sont normées. Soient L et M les dimensions horizontale et verticale du diagramme. Ces quantités sont soit données par l'utilisateur, soit prennent des valeurs muettes. La position du i-ième objet sur l'axe horizontal (l'axe OX) à partir de l'extrémité gauche du diagramme est donnée par la formule

$$n = [L\,(x_i - x_{\min})/(x_{\max} - x_{\min}) + 0{,}5],$$

où [ ] désigne la partie entière, $x_{\max}$ et $x_{\min}$ respectivement les valeurs maximale et minimale de la coordonnée x. On détermine de façon analogue la position d'un objet (le numéro de la ligne) sur la verticale. Au point possédant ces coordonnées on inscrit un symbole quelconque, par exemple, la lettre A. Si k objets tombent en un même point, on inscrit le nombre de ces objets si $k < 10$, et le symbole * si $k \geqslant 10$. S'il existe plusieurs groupes d'objets, on attribue à chacun d'eux un symbole personnel. Les valeurs des coordonnées X et Y sont imprimées dans les échelles correspondantes toutes les cinq positions sur l'axe horizontal et toutes les cinq lignes sur l'axe vertical du diagramme.

12.1.7. Estimation des paramètres et localisation des observations anomales. Considérons le logiciel relatif à l'estimation des paramètres.

1. Estimations standards des principales caractéristiques numériques des variables aléatoires : moyenne, médiane, variance, étendue, coefficient de dissymétrie, coefficient d'aplatissement, etc.

2. Estimations des paramètres d'une loi connue *a priori* : loi gamma, bêta, de Poisson, binomiale, etc.

3. Estimations standards des paramètres de variables aléatoires vectorielles : vecteur des moyennes et matrice des covariances.

4. Estimations stables des paramètres de localisation (cf. n° 10.4.5) :

1) moyenne tronquée et médiane ;
2) moyenne winsorienne ;
3) estimations de Hampel, Tuckey, Andrews ;
4) λ-moments.

5. Estimation stable des paramètres d'échelle.

Le calcul des estimations standards 1 et 3 est conduit à quelques variations près dans chacun des paquets de programmes énumérés. Les estimations des paramètres dans le cas où la loi de probabilité est connue sont réalisées dans le paquet SOD et dans le PSS [88, vyp. 10, razdel 4]. Le logiciel des estimations stables est le moins

développé. Les estimations 4.1 à 4.3 sont réalisées dans les paquets *BMDP* 75 et *BMDP* 79 (programme *P2D*) et dans le PPSA (programme *ESTM*), ainsi que dans le PSS [88, vyp. 2, 10].

Attardons-nous sur la méthode de réalisation des estimations stables du vecteur des moyennes et de la matrice des covariances, qui figure dans le PPSA [108]. Le défaut des estimations pondérées (10.20), (10.21) y compris les estimations des λ-moments (10.20′), (10.21′) réside dans le fait que leurs variances croissent avec la dimension de l'espace des variables p [33]. Pour remédier à cet état de choses, le programme *ESTM* du PPSA prévoit la démarche suivante. Pour chacune des p variables on établit p estimations stables unidimensionnelles de la moyenne et de la variance, puis on résout $p(p-1)/2$ problèmes bidimensionnels pour déterminer les estimations stables des covariances, par exemple, par la méthode des λ-moments. Les difficultés sont de la même nature pour l'estimation de la matrice des covariances en cas de données manquantes, savoir que la matrice des covariances peut ne pas être semi-définie positive. Les propriétés asymptotiques de ces estimations qui sont dites réduites sont indiquées dans [34].

12.2. Calcul des fonctions de répartition et de leurs inverses

Dans ce paragraphe on décrit des méthodes de calcul des valeurs des fonctions de répartition, des points de pourcentage et des inverses des fonctions de répartition des lois les plus couramment utilisées : la loi normale, les lois du χ^2 centrée et non centrée, les lois de Fisher centrée et non centrée, la loi bêta, les approximations des lois limites de certains tests non paramétriques, certaines lois discrètes. La comparaison de la valeur de tel ou tel test aux points de pourcentage de la fonction de répartition correspondante constitue généralement la phase finale d'un test d'hypothèses statistiques. Ceci explique la prolifération des tables des valeurs des fonctions de répartition (cf., par exemple, [1], [27]). Par ailleurs, l'évolution des ordinateurs et du logiciel de la statistique appliquée a institué une nouvelle approche de tabulation des fonctions de répartition et des quantités qui leur sont rattachées. Du temps où les calculs se faisaient à la main, les utilisateurs étaient intéressés à dresser des tables volumineuses susceptibles de délivrer les valeurs des fonctions par une interpolation élémentaire. Mais quand on se sert d'un ordinateur, on n'a aucun intérêt à mémoriser les tables et à composer un programme qui appellerait la table nécessaire et calculerait ensuite la valeur cherchée par une interpolation. Il est préférable de rédiger l'algorithme de calcul des valeurs de la fonction avec la précision requise quand bien même il serait compliqué *).

*) Les algorithmes de calcul des fonctions de répartition en ALGOL sont accessibles dans [45], [101] ainsi que dans [37], [38]. On pourra trouver des programmes en FORTRAN dans [88], [116].

Il importe de signaler particulièrement le cas où le calcul des valeurs de l'inverse de la fonction de répartition *) est utilisé pour la génération des nombres aléatoires obéissant à une loi de répartition donnée dans les méthodes de simulation statistique (cf. § 6.3). Vu que généralement ces nombres sont générés par grandes quantités, il faut mettre en œuvre des algorithmes simples de calcul des inverses des fonctions de répartition **) qui n'occupent pas la machine trop longtemps. Il va de soi que les anciennes tables n'ont pas perdu et ne perdront pas leur actualité dans la mesure où les ordinateurs personnels facilitent considérablement leur mise en œuvre.

Quoi qu'il en soit, ce paragraphe met l'accent sur les algorithmes (développement en séries de fonctions puissances et rationnelles, formules d'approximation de tout type) programmables sur ordinateur.

12.2.1. Loi normale. La fonction de répartition de la loi normale d'espérance mathématique m et de variance σ s'écrit

$$\Phi(x;\ m,\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx. \tag{12.5}$$

Comme $\Phi(x;\ m,\ \sigma) = \Phi\left(\frac{x-m}{\sigma};\ 0,\ 1\right)$, il suffit de savoir calculer la fonction $\Phi(x) = \Phi(x;\ 0,\ 1)$ dite fonction de répartition de la loi normale réduite

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x)\, dx,$$

où $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ est la densité de probabilité de la loi normale. Signalons la propriété suivante de la fonction $\Phi(x)$:

$$\Phi(-x) = 1 - \Phi(x).$$

Pour calculer les valeurs de la fonction $\Phi(x)$, on peut la développer en série entière, en fractions continues ou l'approcher par des polynômes de Tchébychev sur divers intervalles. Citons quelques-unes de ces formules d'approximation utiles en programmation.

*) On appelle *inverse* $\psi(\alpha)$ *d'une fonction de répartition continue* $F(x)$ une fonction dont la valeur en un point arbitraire α de $]0,\ 1[$ se définit comme la solution de l'équation $F(x) = \alpha$.

**) Les formules de ce paragraphe, excepté le cas de la loi normale, ne donnent que des valeurs approchées des inverses des fonctions de répartition. Pour calculer ces valeurs avec une précision suffisante, il faut mettre en œuvre une procédure de recherche d'un zéro d'une fonction continue (cf. par exemple [45]) en prenant des valeurs approchées pour valeurs initiales.

Formules d'approximation pour la fonction de répartition:

$$\Phi(x) = 1 - \varphi(x) \sum_{i=1}^{5} b_i t^i + \varepsilon(x), \tag{12.6}$$

où

$$t = 1/(1 + px), \quad p = 0{,}2316419;$$
$$b_1 = 0{,}319381530, \quad b_2 = -0{,}356563782,$$
$$b_3 = 1{,}781477937, \quad b_4 = -1{,}821255978,$$
$$b_5 = 1{,}330274429;$$
$$|\varepsilon(x)| < 7{,}5 \cdot 10^{-8}.$$

L'approximation suivante est moins précise:

$$\Phi(x) = 1 - \varphi(x) \Sigma a_i t^i + \varepsilon(x);$$
$$t = 1/(1 + px), \quad p = 0{,}33267; \tag{12.7}$$
$$a_1 = 0{,}4361836, \quad a_2 = -0{,}1201676, \quad a_3 = 0{,}9372980.$$

L'erreur $|\varepsilon(x)|$ de cette approximation est $\leqslant 10^{-5}$.

Citons encore une formule d'approximation n'exigeant pas le calcul de la densité $\varphi(x)$:

$$\Phi(x) = 1 - \frac{1}{2}\left(1 + \sum_{i=1}^{4} c_i x^i\right)^{-4} + \varepsilon(x); \tag{12.8}$$

$$c_1 = 0{,}196854, \quad c_2 = 0{,}115194,$$
$$c_3 = 0{,}000344, \quad c_4 = 0{,}019527;$$
$$|\varepsilon(x)| < 2{,}5 \cdot 10^{-4}.$$

On dispose d'une formule d'approximation analogue à (12.8) d'erreur $|\varepsilon(x)| < 1{,}5 \cdot 10^{-7}$ ([1], formule (26.2,19)).

Les formules d'approximation exhibées permettent de calculer les valeurs de $\Phi(x;\ 0,\ 1)$ avec la précision requise.

L'*inverse* de la fonction de répartition $\Phi(x;\ m,\ \sigma)$ de la loi normale sera désigné par $\psi(\alpha;\ m,\ \sigma)$. On a la relation

$$\psi(\alpha;\ m,\ \sigma) = m + \sigma\psi(\alpha),$$

où $\psi(\alpha)$ est l'inverse de la fonction de répartition de la loi normale réduite, soit

$$\Phi(\psi(\alpha);\ 0,\ 1) = \alpha \quad (\alpha \in]0,\ 1[).$$

Etant donné que $\psi(\alpha) + \psi(1 - \alpha) = 0$ pour tout $\alpha \in]0,\ 1[$, il suffit de savoir calculer les valeurs de $\psi(\alpha)$ sur le semi-intervalle $[0{,}5,\ 1[$.

Indiquons les formules d'approximation suivantes acquises par modification des formules (26.2.22) et (26.2.23) de [1] (dans ces deux

formules, $t = \sqrt{-2 \ln (1 - \alpha)}$, $\alpha \in [0{,}5,\ 1[$):

$$\psi(\alpha) = t - \frac{a_0 + a_1 t}{1 + b_1 t + b_2 t^2} + \varepsilon_1(\alpha), \tag{12.9}$$

$$a_0 = 2{,}30753, \quad a_1 = 0{,}27061,$$

$$b_1 = 0{,}99229, \quad b_2 = 0{,}04481;$$

$$\psi(\alpha) = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} + \varepsilon_2(\alpha), \tag{12.10}$$

$$c_0 = 2{,}515517, \quad c_1 = 0{,}802853, \quad c_2 = 0{,}010328,$$

$$d_1 = 1{,}432788, \quad d_2 = 0{,}189269, \quad d_3 = 0{,}001308.$$

L'erreur $|\varepsilon_1| < 3 \cdot 10^{-3}$ et $|\varepsilon_2| < 4{,}5 \cdot 10^{-4}$. Citons un autre développement intéressant de $\psi(\alpha)$:

$$\psi^2(\alpha) = \sum_{i=1}^{\infty} a_i [-\ln (1 - 4(\alpha - 1/2))^2]^i. \tag{12.11}$$

Les quatre premiers coéfficients du développement sont:

$$a_1 = \pi/2, \qquad a_2 = 0{,}37068870 \cdot 10^{-1},$$

$$a_3 = 0{,}83209445 \cdot 10^{-3}, \quad a_4 = -0{,}23232430 \cdot 10^{-3}.$$

Les quatre premiers termes du développement fournissent une précision suffisante pour les applications pratiques dans l'intervalle $0{,}03 < \alpha \leqslant 0{,}97$.

12.2.2. Loi du χ^2. La loi du χ^2 est appliquée dans les problèmes d'estimation et de test d'hypothèses statistiques (cf. chap. 11).

La fonction de répartition de la loi du χ^2 à ν degrés de liberté est définie par la formule

$$F_{\chi^2(\nu)}(x) = \begin{cases} 0 \text{ si } x < 0; \\ C_{\chi^2}(\nu) \int\limits_0^x t^{\nu/2-1} e^{-t/2}\, dt \quad (x \geqslant 0), \end{cases} \tag{12.12}$$

où la constante de normalisation $C_{\chi^2}(\nu) = \left(2^{\nu/2} \Gamma\left(\frac{\nu}{2}\right)\right)^{-1}$.

On a

$$F_{\chi^2(1)}(x) = 2\Phi(\sqrt{x}) - 1,$$

pour $\nu = 1$, et

$$F_{\chi^2(2)}(x) = 1 - e^{-x/2},$$

pour $\nu = 2$.

Dans le cas général, on calcule les valeurs de la fonction de répartition (12.12) en la développant en série. Exhibons deux développements en séries entières (cf. [1], formules (26.4.6) et (26.4.7)) convergents pour tout $x > 0$:

$$F_{\chi^2}(\nu)(x) = C_{\chi^2}(\nu)\, x^{\nu/2} \sum_{n=0}^{\infty} \frac{(-1)^n (x/2)^n}{n!\left(n+\frac{\nu}{2}\right)}; \tag{12.13}$$

$$F_{\chi^2(\nu)}(x) = \\ = C_{\chi^2}(\nu)\, e^{-x/2}\left(1+\sum_{n=1}^{\infty} \frac{(x/2)^n}{\left(1+\frac{\nu}{2}\right)\left(2+\frac{\nu}{2}\right)\ldots\left(n+\frac{\nu}{2}\right)}\right). \tag{12.14}$$

Lorsque $\nu > 2$, les termes de la série (12.14) décroissent plus vite que ceux de la série (12.13) pour tous les x.

Pour les grandes valeurs de x ($x > \nu$), il est préférable de se servir des formules d'approximation basées sur les développements de Pade *):

$$1 - F_{\chi^2(\nu)}x = 2C_{\chi^2}(\nu)\, e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1} \frac{E_n(\nu, x)}{F_n(\nu, x)} + \varepsilon_n(x), \tag{12.15}$$

où

$$E_0(\nu, x) \equiv 1, \quad E_1(\nu, x) = \left(2-\frac{\nu}{2}\right)^{-1}\left(1+\frac{x}{2}\right),$$

$$F_0(\nu, x) \equiv 1, \quad F_1(\nu, x) = \left(2-\frac{\nu}{2}\right)^{-1}\left(2+\frac{x}{2}-\frac{\nu}{2}\right).$$

L'erreur $\varepsilon_n(x) \to 0$ lorsque $x \to \infty$ et est strictement négative pour $x \geqslant \nu$, de sorte que l'on a l'inégalité

$$1 - F_{\chi^2(\nu)}(x) \leqslant 2C_{\chi^2}(\nu)\, e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1} \frac{E_n(\nu, x)}{F_n(\nu, x)} \quad (x \geqslant \nu).$$

Les fonctions $E_n(\nu, x)$ et $F_n(\nu, x)$ vérifient une même relation de récurrence que l'on pourra utiliser pour le calcul d'approximations plus exactes:

$$\left(n+2-\frac{\nu}{2}\right) F_{n+1}(\nu, x) = \left(\frac{x}{2}+2n+2-\frac{\nu}{2}\right) F_n(\nu, x) - nF_{n-1}(\nu, x).$$

Pour $n = 2$, on a

$$\frac{E_2(\nu, x)}{F_2(\nu, x)} = \frac{\left(x/2+4-\frac{\nu}{2}\right) P_1(x) - \left(2-\frac{\nu}{2}\right)}{\left(x/2+4-\frac{\nu}{2}\right) Q_1(x) - \left(2-\frac{\nu}{2}\right)},$$

où $P_1(x) = 1+\frac{x}{2}$, $Q_1(x) = 2+\frac{x}{2}-\frac{\nu}{2}$.

*) Luke Y. *Mathematical Functions and their Approximations*. New York, 1975.

Formules d'approximation pour les grands ν. Lorsque ν est élevé, on a les approximations suivantes qui sont suffisamment précises:

$$F_{\chi^2(\nu)}(x) = \Phi(x_1) + \varepsilon_1, \quad x_1 = \sqrt{2x} - \sqrt{2\nu - 1}; \tag{12.16}$$

$$F_{\chi^2(\nu)}(x) = \Phi(x_2) + \varepsilon_2, \quad x_2 = \frac{(x/\nu)^{1/3} - \left(1 - \frac{2}{9\nu}\right)}{\sqrt{2/9\nu}}. \tag{12.17}$$

La fonction de répartition inverse $F^{-1}_{\chi^2(\nu)}(\alpha)$ peut être approchée à l'aide des formules (12.16) et (12.17):

$$F^{-1}_{\chi^2(\nu)}(\alpha) \approx \frac{1}{2}\{\psi(\alpha) + \sqrt{2\nu - 1}\}^2; \tag{12.18}$$

$$F^{-1}_{\chi^2(\nu)}(\alpha) \approx \nu\left\{1 - \frac{2}{9\nu} + \psi(\alpha)\sqrt{\frac{2}{9\nu}}\right\}. \tag{12.19}$$

12.2.3. Loi bêta. La fonction de répartition $F_{\beta(a,b)}(x)$ de la loi bêta de paramètres $a > 0$, $b > 0$ est définie par la formule

$$F_{\beta(a,b)}(x) = C_\beta(a, b)\int_0^x t^{a-1}(1-t)^{b-1}\,dt \quad (0 \leqslant x \leqslant 1);$$
$$C_\beta(a, b) = 1/B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{12.20}$$

On a la relation

$$F_{\beta(a,b)}(x) = 1 - F_{\beta(b,a)}(1-x). \tag{12.21}$$

Si l'un au moins des paramètres a et b est égal à 1, l'intégrale (12.20) se calcule sans peine sous forme finie. Dans le cas général, pour calculer les valeurs de la fonction $F_{\beta(a,b)}(x)$, il est avantageux de la développer en série de Taylor au voisinage de 0:

$$F_{\beta(a,b)}(x) = C_\beta(a, b)\frac{x^a}{a}\left(1 + a\sum_{n=1}^{\infty}\frac{x^n(1-b)(2-b)\ldots(n-b)}{(n+a)\,n!}\right). \tag{12.22}$$

En vertu de la relation (12.21), on peut toujours faire en sorte que $x \in [0, 1/2]$ et, par suite, accélérer la convergence de la série de (12.22). Les dix premiers termes du développement assurent visiblement une erreur relative $\delta < 10^{-4}$, les quatorze premiers, une erreur relative $\delta < 10^{-5}$ uniformément en a et b.

Préalablement on pourra appliquer la formule de récurrence

$$F_{\beta(a,b)}(x) = C_\beta(a, b)\,x^a(1-x)^{b-1}/a + F_{\beta(a+1,\,b-1)}(x)$$

pour réduire le paramètre b de telle sorte que b soit <1.

La fonction inverse $F^{-1}_{\beta(a,b)}(\alpha)$ est approchée par l'expression ([1], formule (26.5.22)):

$$\tilde{F}^{-1}_{\beta(a,b)}(\alpha) \approx \frac{a}{a + be^{2\omega}}, \tag{12.23}$$

où

$$\begin{cases} w = \dfrac{\psi(\alpha)(h+\lambda)^{1/2}}{h} - \left(\dfrac{1}{2b-1} - \dfrac{1}{2a-1}\right)\left(\lambda + \dfrac{5}{6} - \dfrac{2}{3n}\right); \\ h = 2\left(\dfrac{1}{2a-1} + \dfrac{1}{2b-1}\right)^{-1}, \quad \lambda = \dfrac{\psi^2(\alpha)-3}{6}, \end{cases} \tag{12.24}$$

où $\psi(\alpha)$ est l'inverse de la fonction de répartition de la loi normale réduite. Lorsque l'un des paramètres a ou b est voisin de $1/2$, l'erreur de la formule (12.23) est considérable. Ainsi, $\tilde{F}^{-1}_{\beta(a,b)}(\alpha)$ tend vers 1 pour $a \to 1/2$, et vers 0, pour $b \to 1/2$ quel que soit α.

12.2.4. Loi de Fisher. La fonction de répartition de la loi de Fisher est définie par la formule

$$F_{F(\nu_1,\nu_2)}(x) = \frac{\nu_1^{\frac{1}{2}\nu_2}\nu_2^{\frac{1}{2}\nu_2}}{B\left(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2\right)} \int_0^x t^{\frac{1}{2}(\nu_1-2)}(\nu_2+\nu_1 t)^{-\frac{1}{2}(\nu_1+\nu_2)}\,dt, \tag{12.25}$$

$$x \geqslant 0, \quad \nu_1 > 0, \quad \nu_2 > 0.$$

Cette fonction vérifie la relation

$$F_{F(\nu_1,\nu_2)}(x) = 1 - F_{F(\nu_2,\nu_1)}(1/x). \tag{12.26}$$

Cette relation est identique à la relation (12.21) concernant la fonction de répartition de la loi bêta (compte tenu de (12.27)) et permet de calculer les valeurs de $F_{F(\nu_1,\nu_2)}(x)$ uniquement pour $x \geqslant 1$ (ou $x \leqslant 1$). Un procédé de calcul exact des valeurs de $F_{F(\nu_1,\nu_2)}(x)$ est basé sur le lien de cette fonction avec la fonction de répartition de la loi bêta

$$F_{F(\nu_1,\nu_2)}(x) = 1 - F_{\beta(\nu_1/2,\,\nu_2/2)}(y), \tag{12.27}$$

où $y = \nu_2/(\nu_2 + \nu_1 x)$.

La fonction de répartition inverse $F^{-1}_{F(\nu_1,\nu_2)}(\alpha)$ est approchée à l'aide de (12.23) en tenant compte de la relation liant les fonctions de répartition des lois bêta et de Fisher:

$$F^{-1}_{F(\nu_1,\nu_2)}(\alpha) \approx e^{2w},$$

où w est définie par la formule (12.24) pour $b = \frac{\nu_1}{2}$, $a = \frac{\nu_2}{2}$. Les autres méthodes d'approximation de la fonction $F^{-1}_{F(\nu_1,\nu_2)}(\alpha)$ sont décrites dans ([76], chap. 16) et [27].

12.2.5. Loi t de Student. La fonction de répartition de cette loi est donnée par la formule

$$F_{t(\nu)}(x) = \frac{\nu^{-\frac{1}{2}}}{B\left(\frac{1}{2}, \frac{1}{2}\nu\right)} \int_{-\infty}^{x} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} dt. \tag{12.28}$$

Comme dans le cas de la fonction de répartition de la loi F, pour calculer les valeurs de la fonction (12.28), on peut se servir de la relation qui la lie à la fonction de répartition de la loi bêta:

$$F_{t(\nu)}(x) = 1 - \frac{1}{2} F_{\beta\left(\frac{\nu}{2}, \frac{1}{2}\right)}(y),$$

où

$$y = \left(1 + \frac{x^2}{\nu}\right)^{-1} \quad (x > 0).$$

La fonction de répartition $F_{t(\nu)}$ se calcule sans peine pour $x < 0$ grâce à la symétrie:

$$F_{t(\nu)}(-x) = 1 - F_{t(\nu)}(x).$$

Pour calculer les valeurs de la fonction de répartition inverse $F^{-1}_{t(\nu)}(\alpha)$, on se sert du développement suivant ([1], formule (26.7.5)):

$$F^{-1}_{t(\nu)}(\alpha) = x_\alpha + \frac{g_1(x_\alpha)}{\nu} + \frac{g_2(x_\alpha)}{\nu^2} + \frac{g_3(x_\alpha)}{\nu^3} + \ldots, \tag{12.29}$$

où

$$x_\alpha = \psi(\alpha),$$

$$\begin{cases} g_1(x) = \frac{1}{4}(x^3 + x); \\ g_2(x) = \frac{1}{96}(5x^5 + 16x^3 + 3x); \\ g_3(x) = \frac{1}{384}(3x^7 + 19x^5 + 17x^3 - 15x); \\ g_4(x) = \frac{1}{92\,160}(79x^9 + 776x^7 + 1482x^5 - 1920x^3 - 945x). \end{cases} \tag{12.30}$$

La loi t^2, c'est-à-dire la loi du carré d'une variable aléatoire suivant la loi t, présente souvent dans les applications plus d'intérêt que la loi t. La loi t^2 est un cas particulier de la loi F à 1 et ν degrés de liberté, étudiée au n° 6.1.4.

12.2.6. Lois non centrées. Les fonctions de répartition des lois du χ^2 et F non centrées s'expriment par des séries de fonctions:

$$F_{\chi^2(\nu, \lambda)}(x) = \sum_{i=0}^{\infty} a_i F_{\chi^2(\nu+2i)}(x); \tag{12.31}$$

$$F_{F(\nu_1, \nu_2, \lambda)}(x) = \sum_{i=0}^{\infty} a_i F_{F(\nu_1+2i, \nu_2)}(x), \tag{12.32}$$

où

$$a_i = e^{-\lambda/2} \frac{(\lambda/2)^i}{i!}, \quad \sum_{i=0}^{\infty} a_i = 1.$$

Les fonctions de répartition (12.31) et (12.32) peuvent donc être traitées comme un mélange constitué d'un nombre dénombrable de composantes des fonctions de répartition des lois centrées correspondantes. La méthode la plus directe de calcul des valeurs des fonctions de répartition (12.31) et (12.32) consiste à faire la somme des termes des séries (12.31) et (12.32), calculés avec la précision nécessaire. Ceci étant, l'erreur de calcul $\Delta(k, \varepsilon) \leqslant |\varepsilon| + \delta(k)$, où $|\varepsilon|$ est l'erreur de calcul du terme de la série et $\delta(k)$ l'erreur introduite par la troncature de la série à l'ordre k.

La quantité $\delta(k)$ admet la majoration suivante:

$$\delta(k) = \sum_{i=k}^{\infty} a_i F_{\chi^2(\nu+2i)}(x) \leqslant \sum_{i=k}^{\infty} a_i = 1 - F_{\chi^2(2k)}(\lambda).$$

Cette inégalité nous permet d'estimer sans peine le nombre k_0 de termes de la série (12.31) ou (12.32) suffisant pour que $\delta(k_0) < \delta_0$, où δ_0 est une valeur donnée de l'erreur.

Les fonctions de répartition (12.31) et (12.32) peuvent être exprimées par des formules d'approximation faisant intervenir les fonctions de répartition des lois centrées correspondantes:

$$F_{\chi^2(\nu,\lambda)}(x) \approx F_{\chi^2(\nu^*)}(x^*), \tag{12.33}$$

où

$$x^* = \frac{\nu+2\lambda}{\nu+\lambda}\, x, \quad \nu^* = \frac{(\nu+\lambda)^2}{\nu+2\lambda}$$

et

$$F_{F(\nu_1,\nu_2,\lambda)}(x) \approx F_{F(\nu_1^*,\nu_2)}(x^*), \tag{12.34}$$

où

$$x^* = \frac{\nu_1+\lambda}{\nu_1}\, x, \quad \nu_1^* = \frac{(\nu_1+\lambda)^2}{\nu_1+2\lambda}.$$

Ces approximations peuvent être utilisées en particulier pour le calcul des fonctions de répartition inverses des lois non centrées (cf. formules (12.18), (12.19)).

12.2.7. Approximation des ailes des lois de type ω^2. La détermination des valeurs exactes de la fonction de répartition de la statistique d'un test d'hypothèse d'ajustement et des statistiques de type ω^2 de test de l'hypothèse de normalité donne lieu à des calculs assez fastidieux (cf., par exemple, [92]). Cependant les ailes des fonctions de répartition correspondantes sont justiciables d'une approximation simple dont la précision est satisfaisante pour les applications pratiques. Plus exactement, dans [128] on montre de façon empirique que l'aile droite de la fonction de répartition de la statistique ω^2 du test de Mises-Cramer se laisse bien approcher par l'expression

$$\alpha = P(W^2 \geqslant z) = 0{,}05 \cdot e^{(2{,}70-6z)} + \varepsilon. \tag{12.35}$$

L'erreur ε est telle que

$$|\varepsilon| \leqslant 0{,}002 \text{ pour } \alpha \in [0{,}05,\ 0{,}15] \text{ et respectivement } z \in\ ]0{,}274,\ 0{,}461[;$$

$$|\varepsilon| \leqslant 0{,}001 \text{ pour } \alpha \in [0{,}01,\ 0{,}05] \text{ et respectivement } z \in\ ]0{,}461,\ 0{,}743[.$$

Par exemple, pour $\alpha = 0{,}01$ (resp. $z = 0{,}743$), la formule (12.35) nous donne la valeur 0,0094 et pour $\alpha = 0{,}001$, la valeur 0,0007.

Par une factorisation empirique, on peut approximer les ailes droites des fonctions de répartition des statistiques ω^2 proposées pour le test de la normalité (cf. n° 11.1.7). Ainsi, lorsque la moyenne et la variance sont estimées sur le vu de l'échantillon, on a

$$\alpha = 1 - F(W^2 \leqslant z) \approx 2{,}335 \cdot e^{-30{,}406z}, \tag{12.36}$$

et si l'on estime seulement la moyenne sachant la variance, alors

$$\alpha \approx 1 - F(W^2 \leqslant z) = 2{,}0 \cdot e^{-22{,}21z}. \tag{12.37}$$

Le tableau 12.4 sur lequel sont consignées les valeurs, nominales et calculées par les formules (12.36) et (12.37), du seuil de signification α permet de juger de la précision de ces formules.

Tableau 12.4

Valeur nominale de α	Formule (12.36)	Formule (12.37)	Valeur nominale de α	Formule (12.36)	Formule (12.37)
0,5	—	0,4958	0,1	0,1003	0,1010
0,4	—	0,3961	0,05	0,0506	0,05088
0,3	0,2953	0,2981	0,01	0,107	0,01012
0,2	0,1983	0,1998	0,005	0,00505	0,00500
0,15	0,1494	0,1507	0,001	0,00098	0,00095

12.2.8. Loi normale multidimensionnelle *). Soit Z une variable aléatoire normale p-dimensionnelle. La fonction de répartition $F(X)$ de X se définit comme suit: $F(X) = P(z_1 \leqslant x_1, \ldots, z_p \leqslant x_p)$. La fonction de répartition d'une variable aléatoire normale p-dimensionnelle de vecteur des moyennes M et de matrice des covariances Σ sera désignée par $\Phi_p(X;\ M,\ \Sigma)$. Appelons $\Phi_2(x,\ y;\ \rho)$, où ρ est le coefficient de corrélation, la fonction de répartition de la loi normale réduite à deux dimensions ($m_1 = m_2 = 0$, $\sigma_1 = \sigma_2 = 1$). Pour le calcul de la fonction de répartition à deux

*) L'ouvrage [93] passe en revue les travaux et méthodes concernant le calcul des fonctions de répartition des lois normales multidimensionnelles.

dimensions, on dispose de formules facilement programmables. Citons la formule d'Owen [93]:

$$\Phi_2(x, y; \rho) = \frac{1}{2}\Phi(x) + \frac{1}{2}\Phi(y) - T(x, a_1) - T(y, a_2) \quad (x \geqslant 0,\ y \geqslant 0), \tag{12.38}$$

où

$$T(h, a) = \frac{1}{2\pi}\int_0^a \frac{\exp\left\{-\frac{h}{2}(1+x^2)\right\}}{1+x^2}\,dx;$$

$$a_1 = \frac{y-\rho x}{x\sqrt{1-\rho^2}}, \quad a_2 = \frac{x-\rho y}{y\sqrt{1-\rho^2}}.$$

Pour calculer la fonction $\Phi_2(x, y; \rho)$ dans les autres quadrants, on transforme les variables x et y à l'aide des formules suivantes:

$$\Phi_2(x, y; \rho) = \Phi(x) - \Phi_2(x, -y; \rho),$$

$$\Phi_2(x, y; \rho) = \Phi(y) - \Phi_2(-x, y; \rho),$$

$$\Phi_2(x, y; \rho) = \Phi(x) + \Phi(y) - 1 + \Phi_2(-x, -y; \rho).$$

Dans [93] on trouvera des approximations différentes de la fonction $T(h, a)$.

Dans le cas multidimensionnel, le calcul de la fonction de répartition se ramène à une intégration p-uple de la densité de probabilité. Les méthodes de calcul sont basées soit sur le développement de la fonction de répartition en séries entières multidimensionnelles des coefficients de corrélation [76], soit sur l'abaissement de la dimension de l'intégrale, soit sur la simulation de la variable aléatoire vectorielle correspondante (par la méthode de Monte-Carlo) de fonction de répartition donnée, soit en combinant les deux derniers procédés. Penchons-nous seulement sur la méthode d'abaissement de la dimension de l'intégrale qui ne passe que pour une structure bien définie de la matrice des corrélations.

Soit $\mathbf{R} = (\rho_{ij})$ $(i, j = 1, \ldots, p)$ la matrice des corrélations d'une loi normale multidimensionnelle, et soit

$$\rho_{ij} = \sum_{s=1}^{k} \beta_{is}\beta_{js} \quad (i \neq j,\ i, j = 1, \ldots, p). \tag{12.39}$$

Alors

$$\Phi(X; 0, \mathbf{R}) = \frac{1}{(\sqrt{2\pi})^p\,|\mathbf{R}|^{1/2}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} \exp\left(-\frac{1}{2}Z'\mathbf{R}^{-1}Z\right) dZ$$

se représente par l'intégrale k-uple

$$\Phi_p(X;\ 0,\ \mathbf{R}) = \frac{1}{(\sqrt{2\pi})^k} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \prod_{i=1}^{p} \Phi \times$$

$$\times \left(\frac{x_i - \sum_{j=1}^{k} \beta_{ij} y_j}{\left(1 + \sum_{s=1}^{k} \beta_{is}^2\right)^{1/2}} \right) e^{-\frac{1}{2} \sum_{i=1}^{k} y_i^2} dY.$$

Dans le cas particulier où tous les coefficients de corrélation sont égaux et positifs, on obtient $k = 1$, $\rho_{ij} = b^2$ et

$$\Phi_p(X;\ 0,\ \mathbf{R}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \prod_{i=1}^{p} \Phi\left(\frac{x_i - by}{(1+b^2)^{1/2}}\right) e^{-\frac{y^2}{2}} dy. \qquad (12.40)$$

L'importance de la méthode envisagée s'explique par le fait qu'assez souvent la matrice des corrélations peut être approchée assez exactement par des relations de la forme (12.39) en prenant pour quantités β_{is} ($i = 1, \ldots, p$; $s = 1, \ldots, k$) les composantes des vecteurs propres associés aux plus grandes valeurs propres de la matrice $\mathbf{R}$.

Considérons maintenant quelques résultats relatifs à la probabilité d'un vecteur aléatoire de tomber dans une région de forme spéciale. En statistique, on est souvent conduit à calculer une intégrale normale n-uple étendue à un domaine dans lequel les p composantes de X sont toutes strictement positives (appartiennent au premier quadrant):

$$P(X > 0) = 1 - \Phi_p(0;\ 0,\ \Sigma) = L_p(0,\ \Sigma).$$

Ce problème est certes bien plus facile que le problème général, mais il n'en pose pas moins de grosses difficultés analytiques pour $p > 3$ (cf. [76]). Indiquons les résultats acquis pour $p = 2$ et $p = 3$:

$$L_2(0,\ \mathbf{R}) = \operatorname{Arcsin} \rho/(2\pi) + \frac{1}{4},$$

$$L_3(0,\ \mathbf{R}) = \frac{1}{8} + \frac{1}{4\pi}(\operatorname{Arcsin} \rho_{12} + \operatorname{Arcsin} \rho_{13} + \operatorname{Arcsin} \rho_{14}).$$

De la formule (12.40) il s'ensuit que si tous les ρ sont égaux, alors

$$L_p(0,\ \mathbf{R}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\Phi(-\rho y))^p e^{-\frac{y^2}{2}} dy.$$

Des résultats simples peuvent être établis pour la probabilité $P(A^2,\ \Sigma,\ M_0)$ que le vecteur X tombe dans un ellipsoïde de la

forme

$$(X - M_0)' \Sigma^{-1} (X - M_0) \leqslant a^2,$$

où Σ est la matrice des covariances; M_0 le centre de l'ellipsoïde. Cette probabilité s'exprime au moyen de la fonction de répartition de la loi du χ^2 non centrée (cf. n° 12.2.6):

$$P(A^2, \Sigma, M_0) = F_{\chi^2(p, \lambda)}(a^2);$$
$$\lambda = (M - M_0)' \Sigma^{-1} (M - M_0).$$

12.2.9. Lois discrètes. Citons quelques brèves indications pour le calcul des fonctions de répartition de certaines variables aléatoires discrètes.

Loi de Poisson

$$\sum_{j=0}^{k-1} e^{-\lambda} \frac{\lambda^j}{j!} = 1 - F_{\chi^2(\nu)}(x), \tag{12.41}$$

où $\nu = 2k$; $x = 2\lambda$.

Loi binomiale

$$\sum_{i=m}^{n} C_i^n p^i (1-p)^{n-i} = F_{\beta(m, n-m+1)}(p). \tag{12.42}$$

Loi binomiale négative

$$\sum_{i=m}^{n} C_i^{n+i-1} p^n q^i = F_{\beta(m, n)}(q) \quad (q = 1 - p). \tag{12.43}$$

Loi géométrique

$$\sum_{i=0}^{n} (1-q) q^i = 1 - q^n. \tag{12.44}$$

12.2.10. Calcul de l'espérance mathématique des statistiques de rang. Soient $x_1, \ldots, x_n$ des variables aléatoires indépendantes de fonction de répartition $F(x)$ et de densité de probabilité $f(x)$ continues. En rangeant les observations $x_1, \ldots, x_n$ par ordre de grandeur croissante, on obtient un échantillon ordonné $x_{(1)} \leqslant x_{(2)} \leqslant \leqslant \ldots \leqslant x_{(n)}$ que l'on appelle encore série variationnelle. La statistique $x_{(r)}$ s'appelle statistique de rang r (pour plus de détails sur les statistiques de rang, cf. n° 5.6.4).

La densité de probabilité de la statistique de rang r est;

$$f_{r, n}(z) = \frac{n!}{r!(n-r-1)!} \{F(z)\}^{r-1} \{1 - F(z)\}^{n-r} f(z).$$

Les statistiques de rang et la densité de probabilité $\tau^{-1} f\{(x - \mu)/\tau\}$ sont justiciables des égalités

$$\mathrm{E}x_{(r)} = \mu + g_r \tau, \quad \mathrm{V}x_{(r)} = \tau^2 \sigma_r^{(g)},$$

où g_r et $\sigma_r^{(g)}$ ne dépendent pas de μ et de τ et peuvent être déterminées, par exemple, par une intégration des fonctions respectives (z et z^2) avec la densité $f_{r,n}(z)$. Ceci explique que les résultats sont exhibés pour les statistiques de rang pour $\mu = 0$ et $\tau = 1$.

La quantité $\mathbf{E}x_{(r)}$ joue un rôle important dans les applications des statistiques de rang dans les tests d'hypothèses statistiques (cf. § 11.2). Une méthode de calcul de cette espérance mathématique est basée sur la transformation $v = F(x)$ qui nous ramène à des variables aléatoires indépendantes uniformément réparties sur l'intervalle $]0, 1[$. Cette transformation étant monotone, on a

$$v_{(r)} = F(x_{(r)}) \quad \text{et} \quad x_{(r)} = F^{-1}(v_{(r)}).$$

De la formule de $f_{r,n}(z)$, il résulte que les variables $v_{(r)}$ suivent une loi bêta et

$$\mathbf{E}v_{(r)} = \frac{r}{n+1};$$

$$\mathbf{V}v_{(r)} = \frac{r(n-r+1)}{(n+1)^2(n+2)};$$

$$\mathbf{E}[v_{(r)} - \mathbf{E}v_{(r)}]^3 = \frac{2r(n-2r+1)(n-r+1)}{(n+1)^3(n+2)(n+3)}.$$

D'où la valeur approchée

$$\mathbf{E}x_{(r)} \approx F^{-1}(\mathbf{E}v_{(r)}) = F^{-1}\left(\frac{r}{n+1}\right). \tag{12.45}$$

On obtient une meilleure approximation en développant la fonction $F^{-1}(v_{(r)})$ en série de Taylor au voisinage de $\mathbf{E}v_{(r)} = \alpha_r$. Ainsi, en limitant le développement à l'ordre trois, on obtient

$$\mathbf{E}x_{(r)} = \mathrm{e}_r - \frac{1}{2}(f'(\alpha_r)/(f(\alpha_r))^2)\,\mathbf{V}(v_{(r)}) + \\ + \frac{1}{6(f')^3}(2(f')^2 - f'')\,\mathbf{E}(v_{(r)} - \mathbf{E}v_{(r)})^3 + 0\left(\frac{r}{n^4}\right), \tag{12.46}$$

où $\mathrm{e}_r = F^{-1}(\mathbf{E}v_r) = F^{-1}(\alpha_r) = F^{-1}\left(\frac{r}{n+1}\right)$. On obtiendrait des formules plus exactes en utilisant un plus grand nombre de termes de la série de Taylor et, par conséquent, un plus grand nombre de moments de $v_{(r)}$. Le développement (12.46) est identique à celui de Pearson [76]. En se limitant aux deux premiers termes du développement (12.46), on obtient pour la loi normale

$$\mathbf{E}x_r \approx \Phi^{-1}(\alpha_r) + \frac{\alpha_r\sqrt{\pi}}{\sqrt{2}}\,\mathrm{e}^{\frac{\alpha_r^2}{2}}\,\frac{r(n-r+1)}{(n+1)^2(n+2)}. \tag{12.47}$$

Une autre méthode de calcul des valeurs approchées de $\mathbf{E}x_{(r)}$ consiste à remplacer le second membre de l'égalité (12.45) par

$$F^{-1}\left(\frac{r+a}{n+2a+1}\right).$$

On a montré, en particulier, qu'il faut prendre $a = -3/8$ *) pour les statistiques de rang d'un échantillon prélevé dans une population normale.

Conclusions

1. Le traitement primaire des données est présenté d'une façon assez complète dans le logiciel de la statistique appliquée.

2. Il existe des méthodes efficaces de calcul des fonctions de répartition des lois les plus couramment utilisées en statistique appliquée: les lois du χ^2, de Fisher, bêta, de Student, de Gauss. Des procédures simples ont été élaborées pour inclure ces méthodes dans le logiciel de la statistique appliquée.

---

*) Blom G. *Statistical estimates and transformed beta-variables.* N.Y., Wiley, 1958.

NOTATIONS

Observations primaires

n nombre des objets sondés, nombre d'observations d'une variable aléatoire indépendante;

p nombre de caractères (variables) mesurés sur chaque objet;

$x_i^{(k)}$ valeur de la j-ième variable sur l'objet i;

$$\mathbf{X}_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ x_i^{(p)} \end{pmatrix}$$ vecteur colonne des valeurs des p variables relevées sur l'objet i;

$$\mathbf{X}(t) = \begin{pmatrix} x_1^{(1)}(t), & x_2^{(1)}(t), & \ldots, & x_n^{(1)}(t) \\ x_1^{(2)}(t), & x_2^{(2)}(t), & \ldots, & x_n^{(2)}(t) \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(p)}(t), & x_2^{(p)}(t), & \ldots, & x_n^{(p)}(t) \end{pmatrix}$$ matrice des données initiales relevées à l'«instant» t;

$\rho_{ij}(t)$ caractéristique numérique de la proximité (ou de l'éloignement) de deux objets ou de deux variables de numéros i et j à l'« instant » t;

$$\rho(t) = \begin{pmatrix} \rho_{11}(t)\ \rho_{12}(t) \ldots \rho_{1m}(t) \\ \rho_{21}(t)\ \rho_{22}(t) \ldots \rho_{2m}(t) \\ \cdots \\ \rho_{m1}(t)\ \rho_{m2}(t) \ldots \rho_{mm}(t) \end{pmatrix}$$ matrice des distances des objets (alors $m = n$) ou des variables (alors $m = p$) relevées à l' « instant » t.

Notions de théorie des probabilités

$P\{A\}$ probabilité de l'événement A;

$P\{A \mid B\}$ probabilité conditionnelle de l'événement A sachant que B s'est réalisé;

$P\{AB\}$ probabilité de la réalisation simultanée des événements A et B;

$F_\xi(x)$ fonction de répartition de la variable aléatoire ξ;

$F^{(n)}(x)$ fonction de répartition empirique construite sur le vu de n observations;

$F_{\xi^{(1)}, \ldots, \xi^{(p)}}(x^{(1)}, \ldots, x^{(p)})$ fonction de répartition d'une variable aléatoire vectorielle;

$$f_\xi(x^{(1)}, \ldots, x^{(p)}) = \frac{\partial^{(p)} F_\xi(x^{(1)}, \ldots, x^{(p)})}{\partial x^{(1)} \partial x^{(2)}, \ldots, \partial x^{(p)}}$$

densité de probabilité d'une variable aléatoire vectorielle ξ;

$\mathrm{E}g(\xi) = \int g(x) f_\xi(x)\, dx$ espérance mathématique de la fonction $g(\xi)$ de la variable aléatoire ξ;

$m_k = \mathrm{E}\xi^k$ moment initial d'ordre k de la variable aléatoire ξ;

$m_1 = \mathrm{E}\xi$ moyenne théorique ou espérance mathématique de ξ;

$m_k^{(0)} = \mathrm{E}(\xi - m_1)^{(k)}$ moment centré d'ordre k de la variable aléatoire ξ;

$m_2^{(0)} = \sigma^2 = \mathrm{V}\xi$ variance de ξ;

$\beta_1 = m_3^{(0)} (m_2^{(0)})^{-3/2}$ coefficient de dissymétrie;

$\beta_2 = m_4^{(0)} (m_2^{(0)})^{-2} - 3$ coefficient d'aplatissement;

$x_{\text{méd}}$ médiane;

x_{mod} valeur modale (ou mode) de la variable aléatoire;

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-x^2/2}\, dx$ fonction de répartition d'une variable aléatoire normale (gaussienne) d'espérance mathématique 0 et de variance 1;

$\varphi(x; a, \sigma^2)$ densité d'une variable aléatoire normale d'espérance mathématique a et de variance σ^2;

$N(a, \sigma^2)$ loi normale de moyenne a et de variance σ^2;

$N(M, \Sigma)$ loi normale multidimensionnelle de vecteurs des moyennes M et de matrice des covariances Σ;

$\xi \in N$ la variable ξ suit la loi N;

$\Psi(x)$ fonction inverse de $\Phi(x)$ si le contraire n'est pas spécifié;

$u_q(F)$ quantile d'ordre q d'une variable aléatoire de fonction de répartition F;

$\omega_Q = u_{1-Q/100}$ point de pourcentage 100 Q de la loi;

$\mathrm{cov}(\xi^{(i)}, \xi^{(j)})$, σ_{ij} moment mixte d'ordre deux ou covariance des composantes $\xi^{(i)}$ et $\xi^{(j)}$ de ξ;

$\Sigma = (\sigma_{ij})_{i, j=1, \ldots, p}$ matrice des covariances;

$r_{ij} = \frac{\sigma_{ik}}{(\sigma_{jj}\sigma_{kk})^{1/2}}$ coefficient de corrélation entre les variables aléatoires $\xi^{(j)}$ et $\xi^{(k)}$;

det Σ, Tr (Σ) déterminant et trace de la matrice: mesures de dispersion d'une variable aléatoire vectorielle;

v_ξ coefficient de variance.

Notions de statistique mathématique

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ échantillon ordonné ou série variationnelle;

$\hat{m}_k(n)$, $\hat{m}_k^0(n)$ estimations des moments initial et centré sur le vu d'un échantillon de taille n;

S, $\hat{\Sigma}$ estimation de la matrice des covariances;

$\bar{x}(n)$, $\hat{M}_1(n)$ vecteur des moyennes empiriques d'un échantillon de taille n;

s^2 estimation de la variance, s estimation de l'écart-type;

d_n erreur absolue moyenne;

$L(X_1, X_2, \ldots, X_n)$, $L(X, \theta)$ fonction de vraisemblance;

$\mathbf{I}(\theta, X)$ matrice d'information de Fisher;

H hypothèse statistique;

α seuil de signification ou niveau d'un test;

$1 - \beta$ puissance d'un test;

d_{ij} distance entre des observations vectorielles X_i et X_j;

$\chi^2(k)$ valeur aléatoire suivant une loi du χ^2 à k degrés de liberté, statistique du test du χ^2;

D_n^+, D_n^{-1}, D_n statistique du test de Kolmogorov-Smirnov;

ω_n^2 statistique du test du ω^2;

F_{n_1,n_2} variable aléatoire suivant une loi F à n_1, n_2 degrés de liberté, statistique du test de Fisher.

BIBLIOGRAPHIE

1. Abramowitz M. and Stegun I.A. *Handbook of mathematical functions.* Dover Publications, inc., New York, 1964.

2. Aitchison J., Brown J.A. C. *The Lognormal Distribution*, Cambridge Univ. Press, 1957.

3. Aïvazian S. *Modélisation probabilisto-statistique des rapports distributifs dans la société.* — Dans l'ouvrage: « Modélisation statistique des processus économiques ». Outchenyé zapisky po statistiké. M., Statistika, 1980, t. 37 (en russe).

4. Aïvazian S. *Modélisation des revenus familiaux.* — Ekonomika i matématitcheskié métody, 1970, t. VI, nº 2 (en russe).

5. Aïvazian S. *Comparaison des propriétés optimales des tests de Pearson-Neyman et de Wald.* — Téoria véroiatnosti i eio priménénia, 1959, t. IV, nº 1 (en russe).

6. Aïvazian S. *Test d'hypothèses voisines sur la forme de la densité de probabilité dans le critère séquentiel généralisé.* Téoria véroiatnosti i eio priménénia, 1965, T. X, nº 4 (en russe).

7. Aïvazian S. *Etude statistique des dépendances.* Traduction française. Editions Mir, 1978.

8. Aïvazian S. *Sur la stabilité d'une inférence statistique.* Dans l'ouvrage: «IIe Conférence scientifico-technique nationale « Application de l'analyse statistique multidimensionnelle dans l'économie et l'estimation de la qualité de la production ». Tartou, 1981.

9. Aïvazian S., Bejaéva Z., Starovérov O. *Classification des observations multidimensionnelles.* M., Statistika, 1974.

10. Aïvazian S., Enioukov I., Méchalkine L. *Sur la structure et le contenu d'un paquet de programmes en analyse statistique appliquée.* — Dans l'ouvrage: « Algorithmes et logiciel d'analyse statistique appliquée ». M., Naouka, 1980 (en russe).

11. Aïvazian S. *Statistique mathématique appliquée et problème de la stabilité des inférences statistiques.* — In: Data Analysis and informatics, North-Holland Publ. Comp., 1980.

12. Alimov Yu. *Alternative de la méthode de statistique mathématique.* M., Znanié, 1979 (en russe).

13. Albert A. *Regression and the Moore-Penrose pseudo-inverse.* New York-London, Acad. press, 1972.

14. *Analyse statistique multidimensionnelle dans les recherches socio-économiques.* Outchenyé zapisky po statistiké, 1974, t. XXVI, M., Naouka (en russe).

15. Andersen E. *Discrete statistical models with social science applications.* Lorth Holland, 1980.

16. Anderson T.W. *An introduction to multivariante statistical analysis.* N.Y. Wiley; Lnd., Chapman & Hall, cop. 1958.

17. Baroian O., Rvatchev L., Ivannikov Yu. *Modélisation et prédiction des épidémies de grippe sur le territoire de l'U.R.S.S.* M., 1977 (en russe).

18. Bartholomey D.J. *Stochastic models for social processes.* Wiley, 1970.

19. Beale E.M.L., Little R.J.A. *Missing Values in Multivariate Analysis.* — Journ. Royal. Statist. Soc., 1975, Ser. B, vol. 37.
20. Béliaev Yu. *Méthodes probabilistes du contrôle au hasard.* M., Naouka, 1975 (en russe).
21. Bickel P. *On some robust estimates of location.* — Annals of Math. Statistics. 1965, vol. 36.
22. Bishop Y.M., Fienberg S.E., Holland P.W. *Discrete Multivariate Analysis.* Theory and Practice. Cambridge, NIT-Press, 1975.
23. Blackman K. *An extension of the Kolmogorov distribution.* — Annals of Math. Statistics, 1958, vol. 29.
24. BMDP: *Biomedical Computer Programs.* Ed. W. J. Dixon Univ. of California Press. Berkeley — Los Angeles-London. 1975.
25. BMDP: *Biomedical Computer Programs.* Ed. W.J. Dixon Univ. of California Press, 1979.
26. Bolchev L. *Sur les transformations des variables aléatoires.* Téoria véroiatnostei i eio priménénia, 1959, t. IV, vyp. 2 (en russe).
27. Bolchev L., Smirnov N. *Tables de statistique mathématique.* M., Naouka, 1965 (en russe).
28. Borimetchkov M. *Méthode de prédiction de la distribution des ouvriers et des employés en fonction du montant de leur salaire et du revenu moyen de leurs familles.* — Dans l'ouvrage: « Revenus et demande de la population ». M., Statistika, 1968 (en russe).
29. Borovkov A. *Théorie des probabilités.* M., Naouka, 1976 (en russe).
30. Bouslenko N., Golenko D., Sobol I. et autres. *Méthodes des tests statistiques (méthode de Monte-Carlo)*, M., Fizmatguiz, 1962 (en russe).
31. Box J.E., Cox D.R. *An Analysis of Transformations.* — Journ. Royal. Statist. Soc., 1964, ser. B, vol. 26.
32. Boyarski A. *Cours de démographie.* M., Statistika, 1967 (en russe).
33. Chouryguine A. *Estimation des paramètres de la loi normale au vu d'observations exponentiellement pondérées: théorie asymptotique.* — Dans l'ouvrage: « Algorithmes et logiciel d'analyse statistique appliquée ». M., Naouka, 1980 (en russe).
34. Chouryguine A. *Estimations multidimensionnelles des paramètres d'une loi multidimensionnelle.* — Dans l'ouvrage: « Application de l'analyse statistique multidimensionnelle en économie et dans l'estimation de la qualité de la production ». Tartou, 1981 (en russe).
35. Circulation XXXVI, 1968, vol. 37, n° 3. Supplement I.
36. Cochran W.G. *Sampling techniques.* 2 nd. edn. Wiley, New York, 1963.
37. Cooper B.E. *Algorithm AS2. The normal integral.* — Appl. Statist. 1968, vol. 17, n° 2.
38. Cooper B.E. *Algorithm AS4. An auxiliary function for distribution integrals.* — Appl. Statist. 1968, vol. 17, n° 2.
39. Cox D., Hinkley D. *Theoretical Statistics*, London, 1974.
40. Cramer H. *Mathematical Methods of Statistics*, Stocholm, 1946.
41. David, H.A. *Ordre statistics.* New York (a.o.), cop. 1970.
42. Deev A. *Développements asymptotiques des statistiques de l'analyse discriminante.* — Dans l'ouvrage: « Méthodes statistiques de classification ». M., izd-vo MGU, 1972, vyp. 31 (en russe).
43. Diday E. et collaborateurs. *Optimisation en classification automatique.* T. 1. INRIA. Paris. 1980.
44. Domanski C. *Statystyczne testy nieparametryczne.* Warszawa, 1979.
45. Doubner P. *Calcul des fonctions de répartition directes et inverses*, M., izd-vo MGU, 1971, vyp. 19 (en russe).
46. Durbin J. *Kolmogorov — Smirnow tests when parametres are estimated with applications to test expotentiality and test on spacings.* — Biometrika, 1975, vol. 62.
47. Dwyer P.S. *Some applications of matrix derivatives in multivariate analysis.* — Journ. Amer. Statist. Assoc., 1967, vol. 62.

48. Enukov I. *Méthode de codage des variables non quantitatives.* Dans l'ouvrage: « Algorithmes et logiciel d'analyse statistique appliquée ». M., Naouka, 1980 (en russe).

49. Enukov I., Boulyguine V. *Quelques aspects de l'application pratique de l'analyse discriminante.* — Novosti méditsinskoï tekhniki. M., 1975, vyp. 3 (en russe).

50. Enukov I., Koulakova E. *Codes numériques pour les variables qualitatives en analyse discriminante.* — Dans l'ouvrage: « Analyse statistique multidimensionnelle appliquée ». M., Naouka, 1978 (en russe).

51. Ermakov S., Mikhaïlov G. *Cours de modélisation statistique.* M., Naouka, 1976 (en russe).

52. Fedorov V. *Théorie de l'expérience optimale.* M., Naouka, 1971 (en russe).

53. Fédotov A. *DIAS Système automatique de stockage et traitement statistique des données empiriques (brève description).* Novossibirsk, VTS AN SSSR, 1978, t. 1, preprint (en russe).

54. Feller W. *An Introduction to Probability Theory and its Applications.* Vol. 1, New York, 1957.

55. Fisher R.A. *Theory of statistical estimation.* — Proc. of Cambr. Phil. Soc., 1925, vol. 22.

56. Forsythe G., Moler C. *Computer Solution of Linear Algebraic Systems,* Englewood Cliffs, 1967.

57. Gauss C.F. *Theoria motus corporum coelestium.* Hamburg, 1809.

58. Gauss C.F. *Werke,* bd. 4. Göttingen, 1880.

59. Gnédenko B. *The theory of Probability,* 3 éd., Editions Mir, 1976.

60. Gnédenko B., Béliaev Yu., Soloviev A. *Méthodes mathématiques en théorie de la fiabilité.* Traduction française, Editions Mir, 1972.

61. Gniedenko B.W., Korolyuk W.S. *On the maximum disperancy between two empirical distributions.* — In: Selected Translation in Mathematical Statistical and Probability. Inst. Math. Statist. and Amer. Math. Soc., Providence 1961, vol. 1.

62. Grubbs F. *Sample criteria for testing outlying observations.* — Annals of Math. Statistics, 1950, vol. 21.

63. Grubbs F. *Procedures for detecting outlying observations in samples.* — Technometrics, 1969, vol. 11.

64. Guenoche A., Ihm P. *Analyse en composante principale et analyse discriminante dans le cas de données incomplètes.* — In: Raisonnement et méth. math. archéol. Paris, 1977.

65. Gumbell E. *Statistics of Extremes.* New York, 1962.

66. Haberman S.J. *Analysis of Frequency Data.* Chicago. Univ. of Chicago Press, 1974.

67. Hajek J., Sidak Z. *Theory of rang tests.* Prague. « Academia », 1967.

68. Hald A. *Statistical Theory with Engineering Applications.* New York, 1952.

69. Haldane J.B.S. *The approximate normalisation of a class of frequency-distribution.* — Biometrika, 1937, vol. 29.

70. Hill M.O. *Correspondence analysis a Neglected Multivariate Method.* — Appl. Stat., 1974, vol. 23.

71. Hogg R.V. *Adaptive Robust Procedures: a partial Review and some Suggestions for Future Applications and Theory* — Journ. of Amer. Statist. Assoc., 1974, vol. 69.

72. Huber P.S. *Robust Statistics.* Review. — Annals of Math. Statistics, 1972, vol. 43.

73. Huber P.S. *Robust estimation of a location parameter.* — Annals of Math. Statistics, 1964, vol. 35.

74. Kac Mark. « *Kilka zagadnien stochastycty cznych fizyki i matematyki* ». Warzawa. Panstw. wyd-wo naukowae, 1961.

75. Kamensky V. *Méthodes et modèles de l'analyse des proximités multidimensionnelle non métrique* (revue). — Avtaumatika et télémékhanika, 1977, n° 8 (en russe).

76. Kendall M.G. and Stuart A. *Distribution theory*, vol. 1, London, Griffin, 1965-1966.

77. Kendall M.G. and Stuart A. *Inference and relation ship*, vol. 2, London, Griffen. 1967.

78. Kendall M.G. and Stuart A. *Design and analysis and timeseries*, vol. 3, London. Griffin. 1966.

79. Knuth D.E. *The art of computer programming* (sorting and searching, vol. 3, Addison-Wesley, 1973).

80. Kolmogorov A. *Sur la loi lognormale de répartition des dimensions des particules par désagrégation*. — DAN SSSR, 1941, t. XXXI, n° 2 (en russe).

81. Kolmogorov A. *Notions fondamentales de théorie des probabilités*.

82. Kotioukov V. *Quelques modèles statistiques non standards de prédiction en économétrie*. Novossibirsk, izd. NIIJT; 1977 (en russe).

83. Kourotchkine A. *Propriétés optimales des composantes principales de la matrice des covariances λ-pondérée*. Dans l'ouvrage: « Algorithmes et logiciel d'analyse statistique appliquée ». M., Naouka, 1980 (en russe).

84. Kruskal J.B. *Multidimensional Scaling by Optimizing Goodness of Fit to Nonmetric Hypothesis*. — Psychometrika, 1964, vol. 29, n° 1.

85. Lankaster H.O., Hamdan M.A. *Estimation of the correlation coefficient in contigency tables with possibly nonmetrical characters*. — Psychometrika, 1964, vol. 29.

86. Lébédev V., Sokolov A. *Introduction au système de programmation* OS ES. M., Statistika, 1978 (en russe).

87. Lillicfods H.W. *On the Kolmogorov-Smirnov tests for normality with mean and variance unknown*. — Journ. Amer. Statist. Assoc., 1967, vol. 62.

88. Logiciel du SU des ordinateurs. Minsk, izd. Inst. matématiki AN B.S.S.R., 1973, vyp. 2, t. 2; 1976, vyp. 10, t. 6; 1978, vyp. 14, t. 1, vyp. 16, t. 9; 1980, vyp. 25, t. 1, 2 (en russe).

89. Macrae E-1. Ch. *Matrix Derivatives With an Application to An Adaptive Linear Decision Problem*. — Annals of Statistics, 1974, vol. 2, n° 2.

90. Maïstrov L. *Développement de la notion de « probabilité »*. M., Naouka, 80 (en russe).

91. Maronna R.A. *Robust M-estimators of multivariate location and scatter*. — Annals of Statistics, 1976, vol. 4.

92. Martynov G. *Calcul de la fonction de répartition limite des statistiques des tests de normalité de type* ω^2. — Téoria véroatnostei i eio priménénia, 1976, t. 21 (en russe).

93. Martynov G. *Calcul des fonctions de répartition des lois normales*. — Dans l'ouvrage: « Bilan de la science et de la technique. Théorie des probabilités. Statistique mathématique. Cybernétique théorique ». M., VINITI, 1980, t. 17 (en russe).

94. Martynov G. *Test du* ω^2. M., Naouka, 1978 (en russe).

95. Meshalkin L.D. *Maximum Number of Statisticaly significant associations between risk factors and the frequency of an event in observational studies*. — Biometrika, 1973, vol. 60, n° 1.

96. Meshalkin L.D. *Approximation of multidimensional densities by normal distributions*. — 7th International Biometric Conference, Hannover, 1970.

97. Méchalkine L. *Paramétrisation des lois multidimensionnelles*. — Dans l'ouvrage: « Analyse statistique multidimensionnelle appliquée ». M., Naouka, 1978 (en russe).

98. Méchalkine L. *Affectation de valeurs numériques aux variables qualitatives*. — Dans l'ouvrage: « Problèmes statistiques de gestion Vilnius », 1976, vyp. 14 (en russe).

99. Méchalkine L., Kourotchkine A. *Nouvelle méthode de paramétrisation*

des dépendances régressives. Zapiski naoutchn. séminarov LOMI AN SSSR, 1979, t. 87 (en russe).

100. Méchalkine L., Serdobolsky V. *Erreurs de classification des observations multidimensionnelles.* — Téoria véroatnostéï i eio priménénia, 1978, t. XVIII, n° 4 (en russe).

101. *Méthodes numériques de statistique mathématique. Algorithmes et programmes.* M., izd-vo MGU, 1976, s. 22-29 (en russe).

102. Mitropolsky A. *Technique des calculs statistiques.* M., Fizmatguiz, 1961 (en russe).

103. Morris J.M. and a.o. *Incidence and prediction of ischaemic heart disease in London busmen.* Lancet, 12, 1966.

104. Mosteller F., Tukey J.W. *Data analysis and regression.* A second course in statistics. Eddison-Wesley Publishing Company, 1976.

105. Naylor Th. *Computer Simulation Experiments with Models of Economic Systems.* New York, 1971.

106. OTEKS: *Traitement des tableaux de données empiriques. Paquets de programmes appliqués* (version 2.1). Novossibirsk, 1978 (en russe).

107. Orlov A. *Théorie appliquée des mesures.* Dans l'ouvrage: « Analyse statistique multidimensionnelle appliquée ». M., Naouka, 1978 (en russe).

108. *Paquets de programmes appliqués d'analyse statistique multidimensionnelle appliquée.* M., TsEMI AN SSSR, 1982, preprint (en russe).

109. *Paquets de programmes appliqués de traitement statistique des données.* Kiev, Institut de cybernétique, 1979, preprint (en russe).

110. Pearson E., Sekar C. *The efficiency of statistical tools and a criterion for the rejection of outlying observations.* — Biométrika, 1936, vol. 28.

111. Plackett R.L. *The Analysis of Categorical Data.* London, Griffin, 1974.

112. Prokhorov Yu., Rozanov Yu. *Théorie des probabilités.* M., Naouka, 1967 (en russe).

113. P — STAT: *A User-Oriented Langauge for Statistical Analysis.* Princeton Univ., Princeton, New Jersey.

114. Rao C.R. *Linear statistical inference and its applications Wiley,* New York, 1965.

115. Ryzin. — Edited by J. Van Ryzin. *Classification and Clusterung.* University of Wisconsin at Madison, 1976.

116. Recueil des programmes scientifiques en FORTRAN. M., Statistika, 1974, vyp. 1, 2 (en russe).

117. Scheffé H. *The analysis of variance.* Wiley; Lnd., Chapman & Hall cop. 1959.

118. Schmerling D. et col. *Estimations d'experts. Méthodes et applications* (revue). — Dans l'ouvrage: « Méthodes statistiques d'analyse des estimations d'experts ». M., Naouka, 1977, s. 290-382 (en russe).

119. Seber G. *Linear Regression Analysis.* New York, 1977.

120. Sheppard R.M., Carroll J.D. *Parametric representation of nonlineur data structure.* — In: Multivariate analysis. N.Y., Acad. Press, 1966.

121. Sheppard R.M. *The Analysis of Proximaties: Multidimensional Scaling with an unknown Distance Function.* — Psichometrika, vol. 27, 1962.

122. Smirnov N. *Sur l'estimation du plus grand terme de la série des observations.* DAN SSSR, 1941, t. 33 (en russe).

123. Smirnov N. *Estimation de l'écart entre les courbes empiriques de répartition dans deux échantillons indépendants.* — Bul. Moskovskogo in-ta, séria A, 1939, 2 (en russe).

124. Smoliak S., Titarenko B. *Méthodes stables d'estimation.* M., Statistika, 1980 (en russe).

125. SPSS: Statistical Package for the Social Sciences. Second edition. Mc Graw-Hill Book Company. N.Y., 1975.

126. Starovérov O. *Modèle de déplacement de la population.* M., Naouka, 1979 (en russe).

127. Statistique démographique et causes des décès. Annuaire de statistique sanitaire mondiale, t. 1, VOZ. Genève-Moscou, Méditsina, 1975.

128. Stephens M.A. *Use of the Kolmogorov-Smirnov, Cramer-von Misese and related statistics without extensive tables.* — Journ. Royal Statist. Soc., 1970, vol. 32.

129. Student. *Errors of routine analysis.* — Biometrika, 1927, vol. 19.

130. Térékhina A. *Analyse de la structure des données empiriques par les méthodes de l'analyse des proximités multidimensionnelle.* Thèse de candidat es sciences, M., 1975 (en russe).

131. *Théorie générale de statistique.* Sous la rédaction de A. Boyarsky. M., Izd-vo MGU, 1977 (en russe).

132. Tietjen G., Moore H. *Some Grubb's type statistics for the detection of several outliers.* — Technometrics, 1972, vol. 14.

133. Torgerson W.S. *Multidimensional scaling. Theory and Methods.* — Psychometricka, 1962, vol. 17.

134. Toutoubaline V. *Limites d'applicabilité (les méthodes probabilisto-statistiques et leurs possibilités).* M., Znanié, 1977 (en russe).

135. Truett J., Cornfield J., Kannel W.J. *A multivariate analysis of the risk of coronary heart disease in Frammingham.* — Journ. Chron. Dis., 1967, 19.

136. Tukey J.W. *A survey of sampling from contaminated distributions.* In: « Contributions to Prob. and Statist ». (Ed. Olkin I. et al.) Stantord Univ. Press, 1960.

137. *Typologie de la consommation.* M., Naouka, 1978 (en russe).

138. Van der Waerden. *Statistique mathématique.* Trad. par M-me C. Guinchat, A. Degenne, Paris. Dunod, 1967.

139. Vetter W.J. *Derivatives Operations on Matrices.* — IEEE Trans. Automatic Control, 1970, AC — 15.

140. Wald A. *Sequentiel analysis N.Y. Wiley*; Lnd., Chapman & Hall, 1947.

141. Zacks Sh. *The Theory of Statistical Inference*, New York, 1971.

142. Zagarouïko N., Elkina V., Timerkaev V. *Algorithme de reconstitution des données manquantes dans les tableaux empiriques (algorithme* ZET). Dans l'ouvrage: « Systèmes de calcul numérique ». Novossibirsk, SO AN SSSR, vyp. 61, 1975 (en russe).

INDEX TERMINOLOGIQUE

Ailes d'une loi 191
Algorithme « ZET » 343, 356
Analyse des clusters 32
confluente 57
discriminante 32
factorielle 33
de la loi de probabilité 26
des observations anomales 23, 344-347
des proximités linéaire 293-294, 295
des relations structurales 57
de variance bifactorielle 55
— monofactorielle 55
Approche bayesienne des estimations statistiques 239-242
a priori 66
fréquentiste *a posteriori* 67
modélique *a posteriori* 68
Axiomatique de Kolmogorov 76
Axiome d'existence de la probabilité d'un événement 65, 76

Biais asymptotique 213

Caractéristique opérationnelle 254
Carré moyen complet 239
Censure des échantillons 234
Classification des objets ou des variables 31-32
Codage 270, 274, 357
Coefficient d'aplatissement 120
de corrélation 123
de dissymétrie 119
quantile de différenciation 118
de variation 113
Collectif statistique 67
Composantes principales 287-289, 295
Conditions d'action d'un ensemble statistique 36
Courbes de Pearson 50
Covariance 109
Critère d'autoinformativité 32
d'informativité exogène 32
d'optimisation 26

Densité de probabilité *a posteriori* 239
Développements de Pade 365
Dispersion interquantile 278
Dissymétrie 119, 300-303
Distance de Kolmogorov 291
de Mahalanobis 271, 291
Domaine mesurable 86
Données 353-354
groupées 92
, gestion 354
, groupement 140
, homogénéité 24
, inspection 265
, interprétation 16, 17, 18
, introduction et stockage 265

Données
, listing 265
, pondération 233
, traitement statistique primaire 23
, unification 269
, visualisation 285-294
Dossier d'études 264, 294

Ecart-type 113
Echantillon(s) 17, 90-93, 124
parents 25
stratifiés 94
typiques 94
, censure 234
, étendue 92
, taille 92
Echantillonnage double 258
Echelle catégorisée 267
d'intervalles 268
de mesures 266-271
nominale 267
non catégorisée 267
ordinale 267
quantitative 268
des rapports 269
Efficacité asymptotique 214
Ensemble statistique 35, 36, 39, 44
Erreur absolue moyenne 279
Espace probabilisé continu 74-79
— discret 65-74
des événements élémentaires 61
Espérance mathématique 107
Estimation(s) asymptotiquement sans biais 213
sans biais 202, 243
convergente 202, 243
efficace 204-206
exponentielles pondérées 282
de Humpel 281
par intervalle 214-215, 235-239
du maximum de vraisemblance 215
de la méthode des moindres carrés 226-230
Estimations du minimax 241
des paramètres de localisation et d'échelle 276-285
du point fixe 341
quadratique moyenne 279
robustes 217
superefficace 211
Etude statistique des dépendances 28-31
Evénement(s) aléatoire 61
certain 64
composé 62
contraire 63
élémentaire 61
impossible 64
incompatibles 64
, différence 63
, produit 63
, somme 62
, système complet 64
Expérience aléatoire 60-61

Fichier 17
Fonction de répartition 96
— asymptotique 213
de variables aléatoires 193
de vraisemblance 206
Fonctionnelle de qualité 19, 28, 30, 32
Formule d'addition des probabilités 69
de Bayes 74
de composition 196
de convolution 196
d'Owen 371
des probabilités conditionnelles 71
— totales 73
Fréquence cumulée 96

Histogramme 100, 266, 271-273, 294
Hypothèse (s) statistique 245-248
multiple 250, 254
nulle 248
simple 250

Identificateur abrégé 264
complet 265
Indépendance des événements 71-73
statistique d'une suite d'observations 25
des variables aléatoires 105, 124
Inégalité de Rao-Cramer-Fréchet 209, 243
de Tchébychev 184
Intervalle de confiance 215, 235-239, 243

Loi(s) conjointe 137, 164, 370
— réduite 134
de Pareto 147, 176-177
de Poisson 131-132, 163, 168-169, 212. 219, 224, 373
polynomiale 132-133, 168-169, 254
t de Student 153-154, 178-179, 367-368
tronquée 235
uniforme (rectangulaire) 141-143, 160-161, 172-173, 220
de Weibull 143-147, 174-175

λ-matrice des covariances 283
λ-vecteur moyen 283
Lemme de Neyman-Pearson 252
Listing des données 265
Logarithme de la fonction de vraisemblance 218
Logiciel 350-361
Loi(s) de probabilité 86-90,
bêta 158, 182, 366
binomiale 126-129, 166, 238, 373
— négative 128-129, 162-163, 166-167, 373
de Cauchy 147-148, 178-179
conditionnelle 90, 124
exponentielle 143-147, 164, 172-173. 211, 221, 225
— bilatérale (de Laplace)146-147, 176-177
F de Fisher 154-157, 180-181, 367
gamma 157-158, 180-181, 195
géométrique 373
des grands nombres 186
hypergéométrique 129-131, 166-167
du χ^2 150-153, 178-179, 364-366
λ-liée 283
lognormale 137-141, 170-171, 194
marginale 89, 124
modèle 126
non centrées 157, 368-369
normale 133-137, 164, 168-169
— bidimensionnelle 136-137

Matrice de connexion 85
des corrélations 123
des covariances 109
— w-pondérée 282
d'information de Fisher 207, 209, 238
des observations (du plan) 22, 54, 227
Médiane 111, 267
Mélange de lois 149, 281
Menu 265
Mesure d'informativité 286, 287, 288
probabiliste 77
Méthode(s) d'assistance de l'ordinateur 21
du code moyen 324
du maximum de vraisemblance 215-223, 243
des moindres carrés 226-231, 243
des moments 223-226, 243
d'organisation de l'échantillonnage 93
paramétriques 30
du rang aléatoire 324
des résidus 160
Mode 111, 267
de raisonnement, probabiliste 40
—, probabilistico-statistique 41-43
—, statistique 40

Modèle(s) confluents 67
géométrique 57
markoviens 58
mathématique 20, 28, 39, 40, 45-48, 53-59
de mélange de lois 149
de pollution de Chouryguine 150
de régression 29-31, 54-57
de Tuckey de la loi normale polluée 150, 232, 281, 344
Modélisation mathématique 12, 48-52
statistique 46
Moment(s) centré 108
initial 108
mixtes 109
pondérés 233
Moyenne empirique 110
géométrique 110
harmonique 110
théorique 109
tronquée d'ordre α 280, 360
winsorienne d'ordre α 280, 360

Niveau de confiance 215
d'un test 249
Nombres pseudo-aléatoires 161

Observations extrêmes (anomales) 281, 346, 360
manquantes 24, 337-343, 355-357

Papier gausso-arithmétique 274, 277, 295, 345
Paquets de programmes 350-360
Paramètre de décentrage 157
Paramétrisation de la loi de probabilité de la population générale 26
Période d'exploitation normale 144
de rodage 144
de vieillissement et d'usure 145
Point de pourcentage 117
Pondération des données empiriques 233
Population générale 17, 90-93, 124
— finie 91
— infinie 91
Principe du rapport de vraisemblance 249-251
Probabilité(s) conditionnelle 71, 77
d'un événement 65
objectives 39
Problème de représentativité de l'échantillon 93
Procédure séquentielle d'observation 257
Propriétés asymptotiques des estimations 212-214
préasymptotiques 203
Puissance d'un test 250, 254

Quantile 117-118
Quantité d'information de Fisher 207
Quotient de mortalité 143

Rapport de vraisemblance 241, 251
Recouvrement de l'information 32
Région de confiance 215
Relations de fonctionnement 46
Répartitions w-semblables 282
Représentant d'une population 17
Risque bayesien *a posteriori* 239
de deuxième espèce 248, 254
de première espèce 248, 254

Série de Gram-Charlier-Edgeworth 119
variationnelle 115, 125, 234
Seuil de signification 249
Simulation statistique 52
Stabilité des estimations 278-280, 295
des fréquences relatives 187
des inférences statistiques 27-28
statistique 185, 187-189
Statistique(s) critique 249
descriptive 264-296
de rang 115, 125, 373
— pondérées 232
Suite d'épreuves de Bernoulli 127, 240

Table de survie 36
Tableau de contingence 275-276, 295
Test(s) d'Abbe 335-336
d'ajustement 26
de Bartlett 329
sans biais 255
des blocs 332-334
des codes normaux 318
convergent 255
du E_k 347
de Fraser-Klotz 323
d'homogénéité 310-332
de Klotz 320-321
de Kolmogorov-Smirnov 304-306, 308-310, 348
de Kruskall-Wallis 321
du Λ 330-331
du L_k 347
les plus puissants 251-253
du rapport de vraisemblance 252
séquentiel généralisé 260-261
des séries 334-335
de Student 326
de symétrie 322-324
du T_n 346
du T^2 330, 349
uniformément le plus puissant 252
de Van der Waerden 323
de Wilcoxon-Mann-Whitney 317, 322
du χ^2 298-300, 314-315, 348
du ω^2 306, 308, 309, 348, 369
Théorème de Bernoulli 187
limite central 190
— — multidimensionnel 190
de Moivre-Laplace 192
de Slutsky 188
Tirage aléatoire simple 94
par grappes 94
mécanique 94
séquentiel 95
typique 94
Traitement des coïncidences 324-325
Transformation des variables aléatoires 193-198, 274-275, 295, 355

Unification des données 269
des variables 25

Valeurs empiriques 83
codées 83
groupées 25
individuelles 25
possibles 83
théoriques 83
Variable(s) aléatoire 81, 123
— continue 84
— discrète 84
— endogènes 29
— exogènes 29
— explicatives 29, 48
— nominale 85
— — catégorisée 85
— — non catégorisée 85
— ordinale 85
— prédictantes 29, 55
— quantitative 85
— résultantes 29
— scalaire 84
— vectorielle 84
— complète 337
Variance 113
asymptotique 213
empirique 113
généralisée 122
— empirique 122
intergroupe 330
intragroupe 330
totale 330
Vecteur moyen w-pondéré 282
Visualisation des données 285-294

Zone d'indifférence (d'indétermination) 258